Counterfactuals

William B. Starr
will.starr@cornell.edu

Fall 2018, Final Draft of *SEP* Entry

Modal discourse concerns alternative ways things can be, e.g., what might be true, what isn't true but could have been, what should be done, etc. This entry focuses on **counterfactual modality** which concerns what is not, but could or would have been. What if Martin Luther King had died when he was stabbed in 1958 (Byrne 2005: 1)? What if the Americas had never been colonized? What if I were to put that box over here and this one over there? These modes of thought and speech have been the subject of extensive study in philosophy, linguistics, psychology, artificial intelligence, history and many other allied fields. These diverse investigations are united by the fact that counterfactual modality crops up at the center of foundational questions in these fields.

In philosophy, counterfactual modality has given rise to difficult semantic, epistemological and metaphysical questions:

Semantic How do we communicate and reason about possibilities which are remote from the way things actually are?

Epistemic How can our experience in the actual world justify thought and talk about remote possibilities?

Metaphysical Do these remote possibilities exist independently from the actual world, or are they grounded in things that actually exist?

These questions have attracted significant attention in recent decades, revealing a wealth of puzzles and insights. While other entries address the epistemic — The Epistemology of Modality — and metaphysical questions — Possible Worlds, Actualism — this entry focuses on the semantic question. It will aim to refine this question, explain its central role in certain philosophical debates, and outline the main semantic analyses of counterfactuals.

Section 1 begins with a working definition of counterfactual conditionals (§1.1), and then surveys how counterfactuals feature in theories of agency, mental representation and rationality (§1.2), and how they are used in metaphysical analysis and scientific explanation (§1.3). Section 1.4 then details several ways in which the logic and truth-conditions of counterfactuals are puzzling. This sets the stage for the sections 2 and 3, which survey semantic analyses of counterfactuals that attempt to explain this puzzling behavior.

Section 2 focuses on two related analyses that were primarily developed to study the *logic* of counterfactuals: strict conditional analyses and similarity analyses. These analyses were not originally concerned with saying what the truth-conditions of particular counterfactuals are. Attempts to extend them to that domain, however, have attracted intense criticism. Section 3 surveys more recent analyses that offer more explicit models of when counterfactuals are true. These analyses include premise semantics (§3.1), conditional probability analyses (§3.2) and structural equations/causal models (§3.3). They are more closely connected to work on counterfactuals in psychology, artificial intelligence and the philosophy of science.

While sections 2 and 3 of this entry will require some familiarity with basic set theory and logical semantics, they also provide intuitive characterizations alongside formal definitions. For introductions to these methods see Basic Set Theory — a supplement to Set Theory — and §4 of Classical Logic, as well as Gamut (1991) and Sider (2010). For more mathematically thorough entries discussing counterfactuals see The Logic of Conditionals and Modal Logic.

Contents

1 Counterfactuals and Philosophy

This section begins with some terminological issues (§1.1). It then provides two broad surveys of research that places counterfactuals at the center of key philosophical issues. Section 1.2 covers the role of counterfactuals in theories of rational agency, mental representation and knowledge. Section 1.3 focuses on the central role of counterfactuals in metaphysics and the philosophy of science. Section 1.4 will then bring a bit of drama to the narrative by explaining how counterfactuals are deeply puzzling from the perspective of classical and modal logics alike.

This entry will not cover one prominent use of counterfactuals in philosophy: to explain how philosophical thought experiments could produce knowledge (Williamson 2005, 2007). This is covered at length in §3 of the entry The Epistemology of Modality and in §3.2 of the entry Thought Experiments.

1.1 What are Counterfactuals?

In philosophy and related fields, counterfactuals are taken to be sentences like:

(1) If colonial powers hadn't invaded, the Americas would be very different.

This entry will follow this widely used terminology to avoid confusion. However, this usage also promotes a confusion worth dispelling. Counterfactuals are not really conditionals with contrary-to-fact antecedents. For example (2) can be used as part of an argument that the antecedent is true (Anderson 1951):

(2) If there had been intensive agriculture in the Pre-Columbian Americas, the natural environment would have been impacted in specific ways. That is exactly what we find in many watersheds.

On these grounds, it might be better to speak instead of *subjunctive conditionals*, and reserve the term *counterfactual* for subjunctive conditionals whose antecedent is assumed to be false in the discourse.[1] While slightly more enlightened, this use of the term does not match the use of *counterfactuals* in the sprawling philosophical and interdisciplinary literature surveyed here, and

[1] See Declerck & Reed (2001: 99) and Brée (1982). See also von Fintel (1999) for a closely related definition.

has its own drawbacks that will be discussed shortly. This entry will use *counterfactual conditional* and *subjunctive conditional* interchangeably, hoping to now have dispelled the suggestion that all counterfactuals, in that sense, have contrary-to-fact antecedents.

The terminology of indicative and subjunctive conditionals is also vexed, but it aims to get at a basic contrast which begins between two different forms of conditionals that can differ in truth value. (3) and (4) can differ in truth-value while holding fixed the world they are being evaluated in.[2]

(3) If Oswald didn't kill Kennedy, someone else did. **Indicative**

(4) If Oswald hadn't killed Kennedy, someone else would've. **Subjunctive**

It is easy to imagine a world where (3) is true, and (4) false. Consider a world like ours where Kennedy was assassinated. Further suppose Oswald didn't do it, but some lone fanatic did for deeply idiosyncratic reasons. Then (3) is true and (4) false. Another aspect of the contrast between indicative and subjunctive conditionals is illustrated in (5) and (6).

(5) # Bob never danced. If Bob danced, Leland danced.

(6) Bob never danced. If Bob had danced, Leland would have danced.

(7) Bob never danced. If Bob were to dance, Leland would dance.

Indicatives like (5) are infelicitous when their antecedent has been denied, unlike the subjunctives like (6) and (7) (Stalnaker 1975; Veltman 1986).

The indicative and subjunctive conditionals above differ from each other only in particular details of their linguistic form. It is therefore plausible to explain their contrasting semantic behavior in terms of the semantics of those linguistic differences. Indicatives, like (3) and (5), feature verbs in the simple past tense form, and no modal auxiliary in the consequent. Subjunctives, like (4) and (6), feature verbs in the past perfect (or 'pluperfect') with a modal *would* in the consequent. Something in the neighborhood of these linguistic and semantic differences constitutes the distinction between **indicative and subjunctive conditionals** — summarized in Figure 1.[3]

As with most neighborhoods, there are heated debates about the exact boundaries and the names — especially when future-oriented conditionals are included. These debates are surveyed in the supplement Indicative and Subjunctive Conditionals. The main entry will rely only on the agreed-upon paradigm examples like (3) and (4). The labels *indicative* and *subjunctive* are also flawed since these two kinds of conditionals are not really distinguished on the basis of whether they have indicative or subjunctive mood in the antecedent or conse-

[2] Lewis (1973b: 3) attributes (4) and (3), and this observation, to Adams (1970). But the actual pair in Adams (1970) is *if Oswald hadn't shot Kennedy, Kennedy would be alive today* and *if Oswald didn't shoot Kennedy, Kennedy is alive today*, and the observation made there is that the subjunctive is *justified* while the indicative is not. Adams is careful to say that this does not imply that one is true while the other false.

[3] Where 'V' is a variable ranging over un-tensed verbs.

	Examples	**Antecedents**	**Consequents**	**Deny Antecedent?**
Indicative	(3), (5)	*V-ed*, ...	*V-ed*, ...	Not felicitous
Subjunctive	(4), (6)	*had V-ed*, *were to V*, *V-ed*, ...	*would have V*, *would V*, *would have V-ed*, ...	Can be felicitous

Figure 1: Rough Guide to Indicative and Subjunctive Conditionals

quent.[4] But the terminology is sufficiently entrenched to permit this distortion of linguistic reality.

Much recent work has been devoted to explaining how the semantic differences between indicative and subjunctive conditionals can be derived from their linguistic differences — rather than treating them as semantically unrelated. Much of this work has been done in light of Kratzer's (1986; 2012) general approach to modality according to which all conditionals are treated as two-place modal operators. This approach is also discussed in the supplement Indicative and Subjunctive Conditionals.[5] This entry will focus on the basic logic and truth-conditions of subjunctive conditionals as a whole, and will use the following notation for them (following Stalnaker 1968).[6]

Subjunctive Conditionals (Notation)

> $\phi > \psi$ symbolizes *if it had been the case that ϕ then it would have been the case that ψ*

This project and notation has an important limitation that should be highlighted: it combines the meaning of the modal *would* and *if... then...* into a single connective '>'. This makes it difficult to adequately represent subjunctive

[4] Those labels are used across languages to distinguish two broad functional categories of verbal mood that indicates whether the speaker is committing to the occurrence of the event described by that verb (Palmer 1986: §1.1.2) — much as verbal tense indicates whether that event occurred in the past, present or future. Indicative indicates the clause is being committed to, while the subjunctive is noncommittal (it often includes imperatives, optatives, interrogatives). While *were*-conditionals such as (7) could be said to have an antecedent in the subjunctive mood, the same cannot be said of (4), which is formally indicative past perfect. Further, some languages have a widely used subjunctive mood, but do not employ it in the relevant conditionals (Palmer 1986; Iatridou 2000). Many linguists working on non-Indo European languages use the labels 'realis' and 'irrealis' in a related but different way (Palmer 1986: §§6-7). Stone (1997: 8) suggests terminology implicit in typological work: *remote* and *vivid* modality.

[5] For surveys on indicative conditionals see the complementary entry Indicative Conditionals and Gillies (2012). For a survey of subjunctive conditionals see von Fintel (2012).

[6] This entry will use ϕ, ψ as variables ranging over any sentences of the language, A as a variable ranging over atomic sentences, and $\mathsf{A}, \dots, \mathsf{Z}$ as particular atomic sentences.

conditionals like:

(8) a. If Maya had run, she might have been elected.
b. If Maya had run, she might have been elected and would have been an excellent Senator.
c. "Mr. Taft never asked my advice in the matter, but if he had asked it, I should have emphatically advised him against thus stating publicly his religious belief." (Theodore Roosevelt)
d. If Maya had run, she probably would have won and she might have won big.

Conditionals like (8a) have figured in debates about the semantics of counterfactuals and have been modeled either as a related connective (Lewis 1973b: §1.5) or a normal *would*-subjunctive conditional embedded under *might* (Stalnaker 1981; Stalnaker 1984: Ch.7). But the more complex examples (8b)–(8d) highlight the need for a more refined compositional analysis, like those surveyed in Indicative and Subjunctive Conditionals. So, while this notation will be used in §1.4 and throughout §§2 and 3, it should be regarded as an analytic convenience rather than a defensible assumption.

1.2 Agency, Mind and Rationality

Counterfactuals have played prominent and interconnected roles in theories of rational agency. They have figured prominently in views of what agency and free will amount to, and played important roles in particular theories of mental representation, rational decision making and knowledge. This section will outline these uses of counterfactuals and begin to paint a broader picture of how counterfactuals connect to central philosophical questions.

1.2.1 Agency, Choice and Free Will

A defining feature of agents is that they make choices. Suppose a citizen votes, and in doing so choses to vote for X rather than Y. It is hard to see how this act can be a choice without a corresponding counterfactual being true:

(9) If the citizen had wanted to vote for Y, they could have.

The idea that choice entails the ability to do otherwise has been taken by many philosophers to underwrite our practice of holding agents responsible for their choices — see entries Moral Responsibility and Blame. But understanding the precise meaning of the counterfactual *could have* claim in (9) requires navigating the classic problem of free will: if we live in a universe where the current state of the universe is determined (or near enough) by the prior state of the universe and the physical laws, then it seems like every action of every agent, including their 'choices', are predetermined — see entries Free Will and Causal Determinism. So interpreting this intuitively plausible counterfactual (9) leads quite quickly to a deep philosophical dilemma. One can maintain, with some Incompatibilists, that (9) is a false claim about what's physically possible, and revisit the understanding of agency, choice and responsibility above — see entries Incompatibilist Theories of Free Will and Arguments for Incompatibilism.[7] Alternatively, one can maintain that (9) is a true claim about some non-physical sense of possibility, and explain how that is appropriate to our understanding of choice and responsibility — see entry Compatibilism. It is wrong to construe debates about free will as *just* debates about the meaning of counterfactuals. But, the semantics of counterfactuals can have a substantive impact on delimiting the space of possible solutions, and perhaps even deciding between them. The same is true for research on counterfactual thinking in psychology.

Experiments in social psychology suggest that belief in free will is linked to increased counterfactual thinking (Alquist *et al.* 2015). Further, they have shown that counterfactually reflecting on past events and choices is one significant way humans imbue life experiences with meaning and create a sense of self (Galinsky *et al.* 2005; Heintzelman *et al.* 2013; Kray *et al.* 2010). Incompatibilists might be able to cite this result as an explanation for why so many people believe they have free will. It is a specific form of wishful thinking: it is interwoven with the practices of counterfactual reflection that give our lives meaning. Seto *et al.* (2015) support this idea by showing that variation in subjects' belief in free will predicts how much meaning they derive from relevant instances of counterfactual reflection. This might even be used as part of a pragmatic argument for believing in free will: roughly, belief in free will is so practically important, and our knowledge of the world so incomplete, that it is rational to believe that it exists.[8]

1.2.2 Rationality

Counterfactual reflection is not just used for the 'sentimental' purposes discussed above, but as part of what Byrne (2005) calls *rational imagination.* This capacity is implicated in many philosophical definitions of rational agency. According to the standard model, agency involves intentional action — see entries Agency and Action. While choices are intentional actions, intentional actions are a more general class of actions which, on most views, are in part caused by intentions — see entry Intention. One prominent understanding of intentions is that they are prospective (forward looking) mental states that play a crucial role in planning actions. Byrne (2005, 2016: 138) details psychological evidence showing that counterfactual thinking is central to forming rational intentions. People use counterfactual thinking after particular events to formulate plans that will improve the outcome of their actions in related scenarios. Examples include aviation pilots thinking after a near-accident 'If I had understood the controller's words accurately, I wouldn't have initiated the inappropriate landing attempt', and blackjack players thinking 'If I'd gotten the 2, I would have

[7] Some incompatibilists instead boldly reject the forms of determinism and indeterminism that lead to this conflict. This could seen as providing an alternative account of what is physically possible so as to make (9) true.

[8] See Kant (1781/1787/1987: A533/B560-A558-B586) and Smilansky (2000) for something like this pragmatic view, and Pereboom (2014: 176–8) for criticisms of it.

beaten the dealer'. People who reason in this way show more persistence and improved performance in related tasks, while those who dwell on how things could have been worse, or do not counterfactually reflect at all, show less persistence and no improvement in performance. Finally, human rationality can become disordered when counterfactual thinking goes astray, e.g., in depression, anxiety and schizophrenia (Byrne 2016: 140-3).

This psychological research shows that rational *human* agents *do* learn from the past and plan for the future engaging in counterfactual thinking. Many researchers in artificial intelligence have voiced similar ideas (Ginsburg 1985; Pearl 1995; Costello & McCarthy 1999). But, this view is distinct from a stronger philosophical claim: that the nature of rational agency consists, in part, in the ability to perform counterfactual thinking. Some versions of causal decision theory make precisely this claim, and do so to capture similar patterns of rational behavior. Newcomb's Problem (Nozick 1969) consists of a decision problem which challenges the standard way of articulating the idea that rational agents maximize expected utility, and, according to some philosophers (Stalnaker 1972/1981; Gibbard & Harper 1978), shows that causal or counterfactual reasoning must be included in rational decision procedures — see the entries Causal Decision Theory and Normative Theories of Rational Choice: Expected Utility (§1.1) for more detail and references. In a similar vein, work on belief revision theory explores how a rational agent should revise their beliefs when they are inconsistent with something they have just learned — much like a counterfactual antecedent demands — and uses structures that formally parallel those used in the semantics of counterfactuals (Harper 1975; Gärdenfors 1978, 1982; Levi 1988) — see entries Formal Representations of Belief (§3.4) and The Logic of Conditionals.

1.2.3 Mental Representation, Content and Knowledge

The idea that counterfactual reasoning is central to rational agency has surfaced in another way in cognitive science and artificial intelligence, where encoding counterfactual-supporting relationships has emerged as a major theory of mental representation (Chater *et al.* 2010) — see entries The Computational Theory of Mind and Mental Representation. These disciplines also study how states of mind like belief, desire and intention explain rational agency. But they are not satisfied with just showing that certain states of mind can explain certain choices and actions. They aim to explain *how* those particular states of mind lead to those choices and actions. They do so by characterizing those states of mind in terms of representations, and formulating particular algorithms for using those representations to learn, make choices and perform actions.[9]

Many recent advances in cognitive science and artificial intelligence share a starting point with Bayesian Epistemology: agents must learn and decide what to do despite being uncertain what exactly the world is like, and these processes

[9] As Marr (1982) puts it: an abstract mathematical theory of how cash registers work leaves open what system of numerical representation is used (binary, arabic, roman) and what algorithms are employed to manipulate those representations to perform arithmetic operations.

can be modeled in the probability calculus. On a simple Bayesian approach, an agent represents the world with a probability distribution over binary facts or variables that represent what the world is like. But even for very simple domains the probability calculus does not provide computationally tractable representations and algorithms for implementing Bayesian intelligence. The tools of *Bayesian networks*, *structural equations* and *causal models*, developed by Spirtes *et al.* (1993, 2000) and Pearl (2000, 2009) address this limitation, and also afford simple algorithms for causal and counterfactual reasoning, among other cognitive processes. This framework represents an agent's knowledge in a way that puts counterfactuals and casual connections at the center, and the tools it provides have been influential beyond cognitive science and AI. It has also been applied to topics covered later in this entry: the semantic analyses of counterfactuals (§3.2) and metaphysical dependence, causation and scientific explanation (§1.3). For this reason, it will be useful to describe its basics now, though still focusing on its applications to mental representation. What follows is a simplified version of the accessible introduction in Sloman (2005: Ch.4). For a more thorough introduction, see Pearl (2009: Ch.1).

In a Bayesian framework, probabilities are real numbers between 0 and 1 assigned to propositional variables $A, B, C, \ldots$. These probabilities reflect an agent's subjective credence, e.g., $P(A) = 0.6$ reflects that they think A is slightly more likely than not to be true.[10] At the heart of Bayesian Networks are the concepts of *conditional probability* and two variables being *probabilistically independent*. $P(B \mid A)$ is the credence in B conditional on A being true and is defined as follows:

Definition 1 (Conditional Probability) $P(B \mid A) \coloneqq \dfrac{P(A \wedge B)}{P(B)}$

Conditional probabilities allow one to say when B is probabilistically independent of A: when an agent's credence in B is the same as their credence in B conditional on A and conditional on $\neg A$.

Definition 2 (Probabilistic Independence) B is probabilistically independent of A just in case $P(B) = P(B \mid A) = P(B \mid \neg A)$.

Bayesian networks represent relations of probabilistic dependence. For example, an agent's knowledge about a system containing 8 variables could be represented by the directed acyclic graph and system of structural equations between those variables in Figure 2. While the arrows mark relations of probabilistic dependence, the equations characterize the nature of the dependence, e.g. '$H \coloneqq F \vee G$' means that the value of H is determined by the value of $F \vee G$ (but not vice versa).[11] This significantly reduces the number of values that must be stored.[12]

[10] See Interpretations of Probability (§1) for details about the probability calculus.

[11] Pearl (2009) uses '=' instead of ':=', but this can obscure the fact that this is an asymmetric relation: the left-hand side is determined by the right.

[12] The corresponding joint probability distribution requires storing $2^8 = 256$ probability values — one for each boolean combination of the variables — while this Bayesian network would require only 18 — one conditional probability for each boolean combination of the

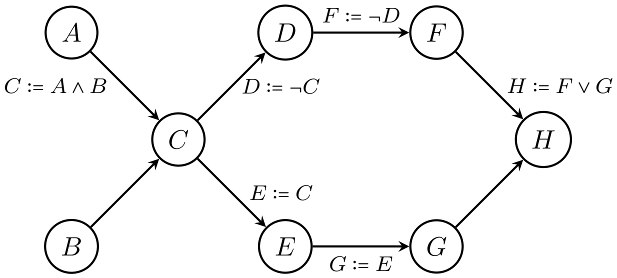

Figure 2: Bayesian Network and Structural Equations

But it also stores information that is useful to agents. It facilitates counterfactual reasoning — e.g. if C had been true then G would have been true — reasoning about actions —- e.g., if we do A then C will be true — and explanatory reasoning — e.g. H is true in part because C is true (Pearl 2002).

The usefulness of Bayesian networks is evidenced by their many applications in psychology (e.g. Glymour 2001; Sloman 2005) and artificial intelligence (e.g. Pearl 2009, 2002). They are among the key representations employed in autonomous vehicles (Thrun *et al.* 2006; Parisien & Thagard 2008), and have been applied to a wide range of cognitive phenomena:

Applications of Bayesian Networks

1. Causal learning and reasoning in AI (Pearl 2009: Chs.1–4) and humans (Glymour 2001; Gopnik *et al.* 2004; Sloman 2005: Chs.6–12)
2. Counterfactual reasoning in AI (Pearl 2009: Ch.7) and humans (Sloman & Lagnado 2005; Sloman 2005; Rips 2010; Lucas & Kemp 2015)
3. Conceptual categorization and action planning (Sloman 2005: Chs.9,10)
4. Learning and cognitive development (Gopnik & Tenenbaum 2007)

As Sloman (2005: 177) highlights, this form of representation fits well with a guiding idea of Embodied Cognition (§3): mental representations in biological agents are constrained by the fact that their primary function is to facilitate successful action despite uncertain information and bounded computational resources. Bayesian networks have also been claimed to address a deep and central issue in artificial intelligence called The Frame Problem (e.g. Glymour 2001: Ch.3). For the purposes of this entry, it is striking how fruitful this approach to mental representation has been, since counterfactual dependence is at its core.

Counterfactual dependence has also featured prominently in theories of mental content, which explain how a mental representation like the concept DOG comes to represent dogs. Informational theories take their inspiration from natural representations like tree rings, which represent, in some sense, how old the

parent variables, and one for each of the two independent variables. See Sloman (2005: Ch.4) and Pearl (2009: Ch.1) for details.

tree is (Dretske 2011). While some accounts in this family are called 'causal theories' — see entry Causal Theories of Mental Content — it is somewhat limiting to formulate the view as: X represents Y just in case Y causes X. Even for the tree rings, it is metaphysically controversial to claim that the tree rings are caused by the age of the tree, rather than thinking they have a common cause or are merely causally related via a number of laws and factors, e.g. rainfall, seasons, growth periods. For this and other reasons, Dretske (1981, 1988, 2002) formulates the relationship in terms of conditional probabilities:

Definition 3 (Dretske's Probabilistic Theory of Information) State s carries the information that a is F, given background conditions g, just in case $P(a \text{ is } F \mid s, g) = 1$.

On this view, the state of the tree rings carries the information that the tree is a certain age, since given the background conditions in our world the relevant conditional probability is 1. As argued by Loewer (1983: 76) and Cohen & Meskin (2006), this formulation introduces problematic issues in how to interpret the probabilities involved and these problems are avoided by a counterfactual formulation:

Definition 4 (Loewer's Counterfactual Theory of Information) State s carries the information that a is F, given background conditions g, just in case, given g, if s were to obtain, a would have to have been F.

Even this theory of information requires several elaborations to furnish a plausible account of mental content. For example, Dretske (1988, 2002) holds that a mental representation r represents that a is F just in case r has the function of indicating that a is F — see entry Teleological Theories of Mental Content. The teleological ('function') component is added to explain how a deer on a dark night can cause tokens of the concept DOG without being part of the information carried by thoughts that token DOG. Fodor (1987, 1990) pursues another, non-teleological solution, the asymmetric dependence theory — see §3.4 of Causal Theories of Mental Content. Counterfactuals feature here in another way:

Definition 5 (Fodor's Asymmetric Dependence Theory) r represents that a is F, just in case:

1. 'a being F causes r' is a law.
2. For any other cause c of r, c would not have caused r if a being F had not caused r. (c's causing r asymmetrically depends on a being F causing r.)

This approach also appeals to laws, which are another key philosophical concept connected to counterfactuals — see §1.3 below.

Counterfactuals are not just used to analyze how a given mental state represents reality, but also when a mental state counts as knowledge. Numerous counterexamples, like Gettier cases, make the identification of knowledge with justified true belief problematic — see entry The Analysis of Knowledge. But some build on this analysis by proposing further conditions to address these counterexamples. Two counterfactual conditions are prominent in this literature (§5.1, 5.2 of The Analysis of Knowledge):

Sensitivity If p were false, S would not believe that p.

Safety If S were to believe that p, p would not be false.

Both concepts are ways of articulating the idea that S's beliefs must be formed in a way that is responsive to p being true. The semantics of counterfactuals have interacted with this project in a number of ways: in establishing their non-equivalence, refining them and adjudicating putative counterexamples. See §§5.1 and 5.2 of The Analysis of Knowledge for details.

1.3 Metaphysical Analysis and Scientific Explanation

Counterfactuals have played an equally central role in metaphysics and the philosophy of science. They have featured in metaphysical theories of causation, supervenience, grounding, ontological dependence and dispositions. They have also featured in issues at the intersection of metaphysics and philosophy of science like laws of nature and scientific explanation. This section will briefly overview these applications, largely linking to related entries that cover these applications in more depth. But, this overview is more than just a list of how counterfactuals have been applied in these areas. It helps identify a cluster of inter-related concepts (and/or properties) that are fruitfully studied together rather than in isolation.

As detailed in Counterfactual Theories of Causation, many philosophers have proposed to analyze causal concepts in terms of counterfactuals (e.g. Lewis 1973a, Mackie 1974). The basic idea is that (10) can be understood in terms of something like (11).

(10) A caused C.

(11) If A had not occurred, C would not have occurred.

This basic idea has been elaborated and developed in several ways. Lewis (1973a, 1979) refines it using his similarity semantics for counterfactuals — see §2.3. The resulting counterfactual analysis of causation faces a number of challenges — see §§3, 4.1 of Counterfactual Theories of Causation. But this has simply inspired a new wave of counterfactual analyses that use different tools.

Hitchcock (2001, 2007) and Woodward (2003: Ch.5) develop counterfactual analyses of causation using the tools of Bayesian networks (or 'causal models') and structural equations described back in §1.2.3. These analyses are covered in §4.2 of Counterfactual Theories of Causation, but the rough idea is easy to state. Given a graph like the one in Figure 2, X can be said to be a cause of Y just in case there is a path from X to Y and changing just the value of X changes the value of Y. According to Hitchcock (2001) and Woodward (2002, 2003), this analysis of causation counts as a counterfactual analysis because the basic structural equations, e.g. $C \coloneqq A \wedge B$, are best understood as primitive counterfactual claims, e.g. if A and B had been true, C would have been true. While not all theories of causation that employ structural equations are counterfactual theories, structural equations are central to many of the contemporary counterfactual theories of causation.[13] See Counterfactual Theories of Causation for further developments and critical reactions to this account of causation.

Recently, Schaffer (2016) and Alastair (2017) have also used structural equations to articulate a counterfactual theory of metaphysical grounding.[14] Metaphysical grounding is a concept widely employed in metaphysics throughout its history, but has been the focus of intense attention only recently — see entry Metaphysical Grounding. As Schaffer (2016) puts it, the fact that Koko the gorilla lives in California is not a fundamental fact because it is grounded in more basic facts about the physical world, perhaps facts about spacetime and certain physical fields. Statements articulating these grounding facts constitute distinct metaphysical explanations. So conceived, metaphysical grounding is among the most central concepts in metaphysics. The key proposals in Schaffer (2016) and Alastair (2017) are to use structural equations to model grounding relations, and not just causal relations, and in doing so capture parallels between causation and grounding. Indeed, they define grounding in terms of structural equations in the same way as the authors above defined causation in terms of structural equations. The key difference is that the equations articulate what grounds what. While this approach to grounding has its critics (e.g. Koslicki 2016), it is worth noting here since it places counterfactuals at the center of metaphysical explanations.[15]

Counterfactuals have been implicated in other key metaphysical debates. Work on dispositions is a prominent example. A glass's fragility is a curious property: the glass has it in virtue of *possibly* shattering in certain conditions, even if those conditions are never manifested in the actual world, unlike say, the glass's shape. This dispositional property — see entry Dispositions — is quite naturally understood in terms of a counterfactual claim:

(12) A glass is fragile if and only if it would break if it were struck in the right way.

Early analyses of this form were pursued by Ryle (1949), Quine (1960) and Goodman (1954), and have remained a major position in the literature on dis-

[13] Spirtes *et al.* (1993, 2000), Pearl (2000, 2009) and Halpern & Pearl (2005a,b) instead treat the equations as representing the 'basic mechanisms' or laws of a causal system. This interpretation is best construed as a non-reductive analysis of causation, rather than analyzing causation in terms of basic counterfactuals. The entry Causation and Manipulability describes how such a view fits into manipulationist theories of causation and the entry Probabilistic Causation describes how it fits into probabilistic theories of causation.

[14] While Alastair (2017) explicitly interprets the structural equations as basic counterfactuals, Schaffer (2016) is less clear on this point. It may be better to read Schaffer (2016) as taking those equations to be basic grounding claims. However, as with causation, there are good reasons to view these equations as counterfactuals (Hitchcock 2001 and Woodward 2002, 2003).

[15] Bennett (2017: §3.3) rejects a counterfactual theory of building relations while taking causation and grounding to be kinds of building relations.

positions. Section 1 of Dispositions overviews the key developments of, and critical reactions to, this view.

It is not just *metaphysical* explanation where counterfactuals have been central. They also feature prominently in accounts of *scientific* explanation and laws of nature. Strict empiricists have attempted to characterize scientific explanation without reliance on counterfactuals, despite the fact that they tend to creep in — see §4 of the entry Scientific Explanation. Scientific explanations appeal to laws of nature, and laws of nature are difficult to separate from counterfactuals. Laws of nature are crucially different from accidental generalizations, but how? One prominent idea is that they 'support counterfactuals'. As Chisholm (1955: 97) observed, the counterfactual (14) follows from the corresponding law (13) but the counterfactual (16) does not follow from the corresponding accidental generalization (15).

(13) All gold is malleable.

(14) If that metal were gold, it would be malleable.

(15) Every Canadian parent of quintuplets in the first half of the 20th century is named 'Dionne'.

(16) If Jones, who is Canadian, had been parent of quintuplets during the first half of the 20th century, he would have been named 'Dionne'.

A number of prominent views have emerged from pursuing this connection. Woodward (2003) argues that the key feature of an explanation is that it answers *what-if-things-had-been-different* questions, and integrates this proposal with a structural equations approach to causation and counterfactuals.[16] Lange (1999, 2000, 2009) proposes an anti-reductionist account of laws according to which they are identified by their invariance under certain counterfactuals. Maudlin (2007: Ch.1) also proposes an anti-reductionist account of laws, but instead uses laws to define the truth-conditions of counterfactuals relevant to physical explanations. For more on these views see §6 of the entry Laws of Nature.

1.4 Semantic Puzzles

It should now be clear that a wide variety of central philosophical topics rely crucially on counterfactuals. This highlights the need to understand their semantics: how can we systematically specify what the world must be like if a given counterfactual is true and capture patterns of valid inference involving them? It turns out to be rather difficult to answer this question using the tools of classical logic, or even modal logic. This section will explain why.

Logical semantics (Frege 1893; Tarski 1936; Carnap 1948) provided many useful analyses of English connectives like *and* and *not* using Boolean truth-functional connectives like $\wedge$ and $\neg$. Unfortunately, such an analysis is not possible for counterfactuals. In truth-functional semantics, the truth of a complex

[16] For Woodward (2003: §5.6), explanations need not involve laws of nature. They only need to involve 'invariants' like the relationships represented in a system of structural equations.

sentence is determined by the truth of its parts because a connective's meaning is modeled as a truth-function — a function from one or more truth-values to another. Many counterfactuals have false antecedents and consequents, but some are true and others false. (17a) is false — given Joplin's critiques of consumerism — and (17b) is true.

(17) a. If Janis Joplin were alive today, she would drive a Mercedes-Benz.
b. If Janis Joplin were alive today, she would metabolize food.

It may be useful to state the issue a bit more precisely.

In truth-functional semantics, the truth-value (True/False: 1/0) of a complex sentence is determined by the truth-values of its parts and particular truth-function expressed by the connective. This is illustrated by the truth-tables for negation $\neg$, conjunction $\wedge$ and the material conditional $\supset$ in Figure 3. Truth-functional logic is inadequate for counterfactuals not just because the material conditional $\supset$ does not capture the fact that some counterfactuals with false antecedents like (17a) are false. It is inadequate because there is, by definition, no truth-functional connective whatsoever that simultaneously combines two false sentences to make a true one like (17b) and combines two false ones to make a false one like (17a). In contemporary philosophy, this is overwhelmingly seen as a failing of classical logic. But there was a time at which it fueled skepticism about whether counterfactuals really make true or false claims about the world at all. Quine (1960: §46, 1982: Ch.3) voices this skepticism and supports it by highlighting puzzling pairs like (18) and (19):

ϕ	$\neg\phi$
1	0
0	1

ϕ	ψ	$\phi \wedge \psi$
1	1	1
1	0	0
0	1	0
0	0	0

ϕ	ψ	$\phi \supset \psi$
1	1	1
1	0	0
0	1	1
0	0	1

Figure 3: Negation ($\neg$), Conjunction ($\wedge$), Material Conditional ($\supset$)

(18) a. If Caesar had been in charge [in Korea], he would have used the atom bomb.
b. If Caesar had been in charge [in Korea], he would have used catapults.

(19) a. If Bizet and Verdi had been compatriots, Bizet would have been Italian.
b. If Bizet and Verdi had been compatriots, Verdi would have been French.

Quine (1982: Ch.3) suggests that no state of the world could settle whether (19a) or (19b) is true. Similarly he contends that it is not the world, but sympathetically discerning the speaker's imagination and purpose in speaking that matters for the truth of (18b) versus (18a) (Quine 1960: §46). Rather than promoting skepticism about a semantic analysis of counterfactuals, Lewis (1973b: 67) took these examples as evidence that their truth-conditions are **context-sensitive**:

the possibilities that are considered when evaluating the antecedent are constrained by the context in which the counterfactual is asserted, including the intentions and practical ends of the speaker. All contemporary accounts of counterfactuals incorporate some version of this idea.[17]

Perhaps the most influential semantic puzzle about counterfactuals was highlighted by Goodman (1947), who noticed that adding more information to the antecedent can actually turn a true counterfactual into a false one. For example, (20a) could be true, while (20b) is false.

(20) a. If I had struck this match, it would have lit.
$\mathsf{S} > \mathsf{L}$
b. If I had struck this match and done so in a room without oxygen, it would have lit.
$(\mathsf{S} \wedge \neg \mathsf{O}) > \mathsf{L}$

Lewis (1973c: 419; 1973b: 10) dramatized the problem by considering sequences such as (21), where adding more information to the antecedent repeatedly flips the truth-value of the counterfactual.

(21) a. If I had shirked my duty, no harm would have ensued.
$\mathsf{I} > \neg \mathsf{H}$
b. Though, if I had shirked my duty and you had too, harm would have ensued.
$(\mathsf{I} \wedge \mathsf{U}) > \mathsf{H}$
c. Yet, if I had shirked my duty, you had shirked your duty and a third person done more than their duty, then no harm would have ensued.
$(\mathsf{I} \wedge \mathsf{U} \wedge \mathsf{T}) > \neg \mathsf{H}$
$\vdots$

The English discourse (21) is clearly consistent: it is nothing like saying *I shirked my duty* and *I did not shirk my duty*. This property of counterfactual antecedents is known by a technical name, *non-montonicity*, and is one of the features all contemporary accounts are designed to capture. As will be discussed in §2.2, even modal logic does not have the resources to capture semantically non-monotonic operators.

Goodman (1947) posed another influential problem. Examples (20a) and (20b) show that the truth-conditions of counterfactuals depend on assumed background facts like the presence of oxygen. However, a moment's reflection reveals that specifying all of these background facts is quite difficult. The match must be dry, oxygen must be present, wind must be below a certain threshold, the friction between the striking surface and the match must be sufficient to produce heat, that heat must be sufficient to activate the chemical energy stored in the match head, etc. Further, counterfactuals like (20a) also rely for their truth on physical laws specific to our world, e.g., the conservation of energy. Goodman's problem is this: it is difficult to adequately specify these background

[17] See Ichikawa (2011), Lewis (2016, 2017b) and Ippolito (2016) for further discussion of the context-sensitivity of counterfactuals.

conditions and laws without further appealing to counterfactuals. This is clearest for laws. As discussed in §1.3, some have aimed to distinguish laws from accidental generalizations by noting that only the former support counterfactuals. But if this is a defining feature of laws, and laws are part of the definition of when a counterfactual is true, circularity becomes a concern. Explicit analyses of laws in terms of counterfactuals, like Lange (2009), would make an analysis of counterfactuals in terms of laws circular.

The potential circularity for background conditions takes a bit more explanation. Suppose one claims to have specified all of the background conditions relevant to the truth of (20a), as in (22a). Then it is tempting to say that (20a) is true because (22c) follows from (22a), (22b) and the physical laws.

(22) a. The match was dry, oxygen was present, wind was below a certain threshold, the potential friction between the striking surface and the match was sufficient to produce heat, that heat was sufficient to activate the chemical energy stored in the match head . . .
b. The match was struck.
c. The match lit.

But now suppose there is an agent seeing to it that a fire is not started, and will only strike the match if it is wet. In this case the counterfactual (20a) is intuitively false. However, unless one adds the counterfactual, *if the match were struck, it would have to be wet*, to the background conditions, (22c) still follows from (22a), (22b) and the physical laws. That would incorrectly predict the counterfactual to be true. In short, it seems that the background conditions must themselves consist of counterfactuals. Any analysis of counterfactuals that captures their sensitivity to background facts must either eliminate these appeals to counterfactuals, or show how this appeal is non-circular, e.g., part of a recursive, non-reductive analysis.

To summarize, this section has identified three key theses about the semantics of counterfactuals and a central problem:

Key Semantic Theses about Counterfactuals

1. Counterfactuals are not truth-functional.
2. Counterfactuals have context-sensitive truth-conditions.
3. Counterfactual antecedents are interpreted non-monotonically.

Goodman's Problem The truth-conditions of counterfactuals depend on background facts and laws. It is challenging to specify these facts and laws in general, but particularly difficult to specify them in non-counterfactual terms.

These theses, along with Goodman's Problem, were once grounds for skepticism about the coherence of counterfactual discourse. But with advances in semantics and pragmatics, they have instead become the central features of counterfactuals that contemporary analyses aim to capture.

2 The Logic of Counterfactuals

This section will survey two semantic analyses of counterfactuals: the **strict conditional** analysis and the **similarity** analysis. These conceptually related analyses also have a shared explanatory goal: to capture logically valid inferences involving counterfactuals, while treating them non-truth-functionally, leaving room for their context dependence, and addressing the non-monontonic interpretation of counterfactual antecedents. Crucially, these analyses abstract away Goodman's Problem because they are not primarily concerned with the truth-conditions of particular counterfactuals — just as classical logic does not take a stand on which atomic sentences are actually true. Instead, they say only enough about truth-conditions to settle matters of logic, e.g. if ϕ and $\phi > \psi$ are true, then ψ is true. Sections 2.5 and 2.6 will revisit questions about the truth-conditions of particular counterfactuals, Goodman's Problem and the philosophical projects surveyed in §1.

The following subsections will detail strict conditional and similarity analyses. But it is useful at the outset to consider simplified versions of these two analyses alongside each other. This will clarify their key differences and similarities. Both analyses are also stated in the framework of possible world semantics developed in Kripke (1963) for modal logics. The following subsection provides this background and an overview of the two analyses.

2.1 Introducing Strict and Similarity Analyses

The two key concepts in possible worlds semantics are possible worlds and accessibility spheres (or relations). Intuitively, a possible world w is simply a way the world could be or could have been. Formally, they are treated as primitive points in the set of all possible worlds W. But their crucial role comes in assigning truth-conditions to sentences: a sentence ϕ can only said to be true given a possible world w, but since w is genuinely possible, it cannot be the case that both ϕ and $\neg\phi$ are true at w. Accessibility spheres provide additional structure for reasoning about what's possible: for each world w, $R(w)$ is the set of worlds accessible from w.[18] This captures the intuitive idea that given a possible world w, a certain range of other worlds $R(w)$ are possible, in a variety of senses. $R_1(w)$ might specify what's nomologically possible in w by including only worlds where w's natural laws hold, while $R_2(w)$ specifies what's metaphysically possible in w.

These tools furnish truth-conditions for a formal language including non-truth-functional necessity ($\Box$) and possibility ($\Diamond$) operators:[19]

[18] Rather than a *sphere* of accessibility, Kripke (1963) uses an accessibility relation $R(w, w')$. Accessibility spheres will fit more smoothly with the presentation here and can be defined in terms of an accessibility relation: $R(w) \coloneqq \{w' \mid R(w, w')\}$.

[19] Here, v is an atomic valuation which assigns every atomic sentence to exactly one truth-value in each possible world. Atomic valuations correspond to one line in a truth-table in classical logic.

Definition 6 (Kripkean Semantics)

1. $[\![A]\!]_v^R = \{w \mid v(w, A) = 1\}$
2. $[\![\neg\phi]\!]_v^R = W - [\![\phi]\!]_v^R$
3. $[\![\phi \wedge \psi]\!]_v^R = [\![\phi]\!]_v^R \cap [\![\psi]\!]_v^R$
4. $[\![\phi \vee \psi]\!]_v^R = [\![\phi]\!]_v^R \cup [\![\psi]\!]_v^R$
5. $[\![\phi \supset \psi]\!]_v^R = (W - [\![\phi]\!]_v^R) \cup [\![\psi]\!]_v^R$
6. $[\![\Box\phi]\!]_v^R = \{w \mid R(w) \subseteq [\![\phi]\!]_v^R\}$
7. $[\![\Diamond\phi]\!]_v^R = \{w \mid R(w) \cap [\![\phi]\!]_v^R \neq \varnothing\}$

In classical logic, the meaning of ϕ is simply its truth-value. But in modal logic, it is the set of possible worlds where ϕ is true: $[\![\phi]\!]$. So ϕ is true in w, relative to v and R, just in case $w \in [\![\phi]\!]_v^R$:

Definition 7 (Truth) $w, v, R \vDash \phi \iff w \in [\![\phi]\!]_v^R$

Only clauses 6 and 7 rely crucially on this richer notion of meaning. $\Box\phi$ says that in all accessible worlds $R(w)$, ϕ is true. $\Diamond\phi$ says that there are some accessible worlds where ϕ is true. Logical concepts like consequence are also defined in terms of relations between sets of possible worlds. The intersection of the premises must be a subset of the conclusion (i.e. every world where the premises are true, the conclusion is true):

Definition 8 (Logical Consequence)
$\phi_1, \ldots, \phi_n \vDash \psi \iff \forall R, v\colon ([\![\phi_1]\!]_v^R \cap \cdots \cap [\![\phi_n]\!]_v^R) \subseteq [\![\psi]\!]_v^R$

Given this framework, the strict analysis can be formulated very simply: $\phi > \psi$ should be analyzed as $\Box(\phi \supset \psi)$. This says that all accessible ϕ-worlds are ψ-worlds. This analysis can be depicted as in Figure 4.[20]

The red circle delimits the worlds accessible from w_0, the x-axis divides ϕ and $\neg\phi$-worlds, and the y-axis ψ and $\neg\psi$-worlds. $\Box(\phi \supset \psi)$ says that there are no worlds in the blue shaded region.

It is crucial to highlight that this semantics *does not* capture the non-monotonic interpretation of counterfactual antecedents. For example, $[\![\mathsf{A} \wedge \mathsf{B}]\!]_v^R$ is a subset of $[\![\mathsf{A}]\!]$, and this means that any time $\Box(\mathsf{A} \supset \mathsf{C})$ is true, so is $\Box((\mathsf{A} \wedge \mathsf{B}) \supset \mathsf{C})$. After all, if all A-worlds are in the red quadrant of Figure 4, so are all of the $\mathsf{A} \wedge \mathsf{B}$-worlds, since the $\mathsf{A} \wedge \mathsf{B}$-worlds are just a subset of the A-worlds. A crucial point here is that on this semantics the domain of worlds quantified over by a counterfactual is constant across counterfactuals with different antecedents. As will be discussed in §2.2, advocates of strict conditional analyses aim to instead capture the non-monotonic behavior of antecedents pragmatically by incorporating it into a model of their context-sensitivity. The

[20] I would like to thank Gabriel Greenberg for allowing me to use this (modified) version of his diagram from Greenberg (2013).

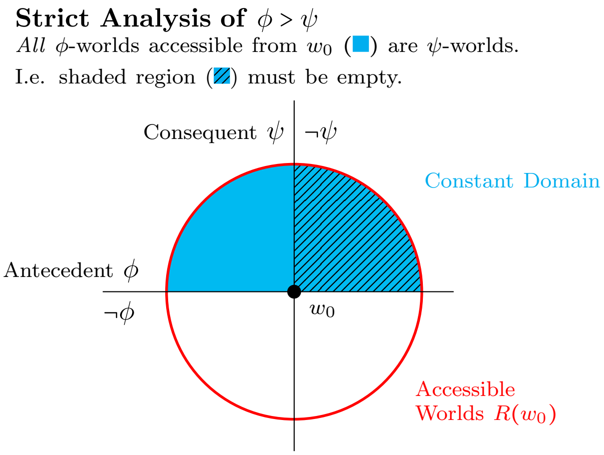
Strict Analysis of $\phi > \psi$
All ϕ-worlds accessible from w_0 (🟦) are ψ-worlds.
I.e. shaded region (▨) must be empty.
Consequent ψ
$\neg\psi$
Constant Domain
Antecedent ϕ
$\neg\phi$
w_0
Accessible Worlds $R(w_0)$

Figure 4: Truth in w_0 relative to R

most important difference between strict analyses and similarity analyses is that similarity analyses capture this non-monotonicity semantically.

On the similarity analysis, $\phi > \psi$ is true in w_0, roughly, just in case all the ϕ-worlds most similar to w_0 are ψ-worlds. To model this notion of similarity, one needs more than a simple accessibility sphere. One way to capture it is with with a nested system of spheres $\mathcal{R}$ around a possible world w_0 (Lewis 1973b: §1.3) — this is just a particular kind of set of accessibility spheres. As one goes out in the system, one gets to less and less similar worlds. This analysis can be depicted as in Figure 5.[21] The most similar ϕ-worlds are in the innermost gray region. So, this analysis excludes any worlds from being in the *shaded* innermost blue region. Comparing Figures 4 and 5, one difference stands out: the similarity analyses does not require that there be no $\phi \wedge \neg\psi$-worlds in *any* sphere, just in the innermost sphere. For example, world w_1 does not prevent the counterfactual $\phi > \psi$ from being true. It is not in the ϕ-sphere most similar to w. This is the key to semantically capturing the non-monotonic interpretation of antecedents. The truth of $\mathsf{A} > \mathsf{C}$ does not guarantee the truth of $(\mathsf{A} \wedge \mathsf{B}) > \mathsf{C}$ precisely because the most similar A-worlds may be in the innermost sphere, and the most similar $\mathsf{A} \wedge \mathsf{B}$ may be in an intermediate sphere, and include worlds like w_1 where the consequent is false. In this sense, the domain of worlds quantified over by a similarity-based counterfactual varies across counterfactuals with different antecedents, though it does express a strict conditional over this varying domain. For this reason, Lewis (1973b) and many others call the similarity analysis a *variably-strict* analysis.

Since antecedent monotonicity is the key division between strict and similarity analyses, it is worthwhile being a bit more precise about what it is, and

[21] I would like to thank Gabriel Greenberg for allowing me to use this (modified) version of his diagram from Greenberg (2013).

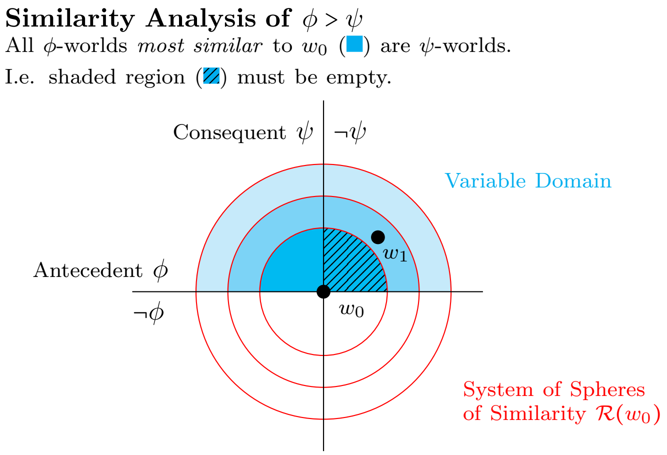
Similarity Analysis of $\phi > \psi$
All ϕ-worlds *most similar* to w_0 (🟦) are ψ-worlds.
I.e. shaded region (▨) must be empty.
Consequent ψ
$\neg\psi$
Variable Domain
Antecedent ϕ
$\neg\phi$
w_1
w_0
System of Spheres
of Similarity $\mathcal{R}(w_0)$

Figure 5: Truth in w_0 relative to $\mathcal{R}$

what its associated inference patterns are.

Definition 9 (Antecedent Monotonicity) If $\phi_1 > \psi$ is true at some w, R, v and $\llbracket\phi_2\rrbracket^R_v \subseteq \llbracket\phi_1\rrbracket^R_v$, then $\phi_2 > \psi$ is true at w, R, v.

The crucial patterns associated with antecedent montonicity are:

Antecedent Strengthening (AS) $\phi_1 > \psi \vDash (\phi_1 \wedge \phi_2) > \psi$

Simplification of Disjunctive Antecedents (SDA)
$(\phi_1 \vee \phi_2) > \psi \vDash (\phi_1 > \psi) \wedge (\phi_2 > \psi)$

Transitivity $\phi_2 > \phi_1, \phi_1 > \psi \vDash \phi_2 > \psi$

Contraposition $\phi > \psi \dashv\vDash \neg\psi > \neg\phi$

AS and SDA clearly follow from antecedent montonicity. By contrast, Transitivity and a plausible auxiliary assumption entail antecedent montonicity,[22] and the same is true for Contraposition.[23] With these basics in place, it is possible to focus in on each of these analyses in more detail. In doing so, it will become clear that there are important differences even among variants of the similarity analysis and variants of the strict analysis. This entry will focus on what these analyses predict about valid inferences involving counterfactuals.

[22] Auxiliary assumption: if $\llbracket\phi_2\rrbracket^R_v \subseteq \llbracket\phi_1\rrbracket^R_v$ then $\phi_2 > \phi_1$ is true at any w in R, v. Suppose that the antecedent of antecedent montonicity holds so (a) $\phi_1 > \psi$ is true at some w, R, v and (b) $\llbracket\phi_2\rrbracket^R_v \subseteq \llbracket\phi_1\rrbracket^R_v$. From (b) and the auxiliary assumption it follows that $\phi_2 > \phi_1$ is true at w, R, v. By Transitivity, $\phi_2 > \psi$ follows.

[23] Auxiliary assumption: if $\phi > \psi_1$ is true and $\llbracket\psi_1\rrbracket^R_v \subseteq \llbracket\psi_2\rrbracket^R_v$, then $\phi > \psi_2$ is true. Suppose that the antecedent of antecedent montonicity holds so (a) $\phi_1 > \psi$ is true at some w, R, v and (b) $\llbracket\phi_2\rrbracket^R_v \subseteq \llbracket\phi_1\rrbracket^R_v$. $\neg\psi > \neg\phi_1$ follows from (a) by Contraposition. From (b) it follows that $\llbracket\neg\phi_1\rrbracket^R_v \subseteq \llbracket\neg\phi_2\rrbracket^R_v$, since $W - \llbracket\phi_1\rrbracket^R_v \subseteq W - \llbracket\phi_2\rrbracket^R_v$ follows from (b) and the set-theoretic fact that $A \subseteq B \iff (W - B) \subseteq (W - A)$. From this and the auxiliary assumption $\neg\psi > \neg\phi_2$ follows. By Contraposition again, $\phi_2 > \psi$ follows.

2.2 Strict Conditional Analyses

The strict conditional analysis has a long history, but its contemporary form was first articulated by Peirce:[24]

> "If A is true then B is true"... is expressed by saying, "In any possible state of things, $[w]$, either $[A]$ is not true [in w], or $[B]$ is true [in w]." (Peirce 1896: 33)

C.I. Lewis (1912, 1914) defended the strict conditional analysis of subjunctives and developed an axiomatic system for studying their logic, but offered no semantics. A precise model-theoretic semantics for the strict conditional was first presented in Carnap (1956: Ch.5). However, that account did not appeal to accessibility relations, and ranged only over logically possible worlds. Since counterfactuals are often non-logical, it it was only after Kripke (1963) introduced a semantics for modal logic featuring an accessibility relation, that the modern form of the strict analysis was precisely formulated:[25]

Basic Strict Conditional Analysis

- $\llbracket \phi > \psi \rrbracket^R_v = \llbracket \Box(\phi \supset \psi) \rrbracket^R_v$
- $\Box(\phi \supset \psi)$ is true in w, relative to R and v, just in case $\phi \supset \psi$ is true at all worlds accessible from w, namely all worlds in $R(w)$
    - I.e. all ϕ-worlds in $R(w)$ are ψ-worlds
- $\llbracket \Box(\phi \supset \psi) \rrbracket^R = \{w \mid R(w) \subseteq \llbracket \phi \supset \psi \rrbracket^R_v\}$
  $= \{w \mid (R(w) \cap \llbracket \phi \rrbracket^R_v) \subseteq \llbracket \psi \rrbracket^R_v\}$
- $\phi \mathrel{-\!\!3} \psi := \Box(\phi \supset \psi)$

Just as the logic of $\Box$ will vary with constraints that can be placed on R, so too will the logic of strict conditionals.[26] For example, if one does not assume that $w \in R(w)$ then modus ponens will not hold for the strict conditional: ψ will not follow from ϕ and $\Box(\phi \supset \psi)$. But even without settling these constraints, some basic logical properties of the analysis can be established. The discussion

[24] Peirce (1896: 33) attributes this view to Philo the Logician, a member of the early Hellenistic Dialectical School. However Bobzien (2011: §3.1) presents Philo as a material implication theorist. For these historical issues see Sanford (1989: Ch.2), Copeland (2002) and Zeman (1997).

[25] Saying that $\phi \supset \psi$ is true throughout $R(w)$ is equivalent to saying that ψ is true throughout the ϕ-worlds in $R(w)$. More formally: (a) $R(w) \subseteq ((W - \llbracket \phi \rrbracket^R_v) \cup \llbracket \psi \rrbracket^R_v)$ holds if and only if (b) $(R(w) \cap \llbracket \phi \rrbracket^R_v) \subseteq \llbracket \psi \rrbracket^R_v$. Suppose (a) and that $w' \in R(w) \cap \llbracket \phi \rrbracket^R_v$. Then $w' \in R(w)$ and by (a) $w' \in ((W - \llbracket \phi \rrbracket^R_v) \cup \llbracket \psi \rrbracket^R_v)$. This entails $w' \in \llbracket \psi \rrbracket^R_v$, after all $w' \in \llbracket \phi \rrbracket^R_v$ in which case $w' \notin (W - \llbracket \phi \rrbracket^R_v)$. Thus (b) follows from (a). Now suppose (b) and that $w' \in R(w)$. Either $w' \in \llbracket \phi \rrbracket^R_v$ or $w' \notin \llbracket \phi \rrbracket^R_v$. Suppose the former. Then $w' \in R(w) \cap \llbracket \phi \rrbracket^R_v$. So by (b) $w' \in \llbracket \psi \rrbracket^R_v$ and so $w' \in ((W - \llbracket \phi \rrbracket^R_v) \cup \llbracket \psi \rrbracket^R_v)$. Suppose the latter. Then $w' \in (W - \llbracket \phi \rrbracket^R_v)$ and so $w' \in ((W - \llbracket \phi \rrbracket^R_v) \cup \llbracket \psi \rrbracket^R_v)$. Thus (a) follows from (b).

[26] For this see the entry Modal Logic.

to follow is by no means exhaustive.[27] Instead, it will highlight the logical patterns which are central to the debates between competing analyses.

The core idea of the basic strict analysis leads to the following validities.

Fact 1 (‘Paradoxes’ of Strict Implication)

1. $\vDash (\phi \wedge \neg\phi) \strictif \psi$

2. $\vDash \psi \strictif (\phi \vee \neg\phi)$

3. $\neg\Diamond\phi \vDash \phi \strictif \psi$

4. $\Box\psi \vDash \phi \strictif \psi$

In these validities, some see a plausible and attractive logic (Lewis 1912, 1914). Others see them as “so utterly devoid of rationality [as to be] a *reductio ad absurdum* of any view which involves them” (Nelson 1933: 271), earning them the title *paradoxes of strict implication*. Patterns 3 and 4 are more central to debates about counterfactuals, so they will be the focus here. Pattern 3 clearly follows from the core idea of the basic strict analysis: the premise guarantees that there are no accessible ϕ-worlds, from which it vacuously follows that all accessible ϕ-worlds are ψ-worlds. Much the same is true of pattern 4: if all the accessible worlds are ψ-worlds then all the accessible ϕ-worlds are ψ-worlds. Both 3 and 4 are seem incorrect for English counterfactuals.

(23) a. JFK couldn’t have passed universal healthcare.
b. If JFK had passed universal healthcare, he would have granted insects coverage.

Contrary to pattern 3, the false (23b) does not intuitively follow from the true (23a). Similarly, for pattern 4. Suppose one’s origin from a particular sperm and egg is an essential feature of oneself. Then (24a) is true.

(24) a. Joplin had to have come from the particular sperm and egg she in fact came from.
b. If there had been no life on Earth, then Joplin would have come from the particular sperm and egg she in fact came from.

And, yet, many would hesitate to infer (24b) on the basis of (24a). Each of these patterns follow from the core idea of the strict analysis. While these counterexamples may not constitute a conclusive objection, they do present a problem for the basic strict analysis. The second wave strict analyses surveyed in §2.2.1 are designed to solve it, however. They are also designed to address another suite of validities that are even more problematic.

The strict analysis is widely criticized for validating antecedent monotonic patterns. It is worth saying a bit more precisely, using Definition 9 and Figure 6, why antecedent monotonicity holds for the strict conditional. If $\phi_1 \strictif \psi$ is

[27] For a more exhaustive study of the logic of strict conditionals see Cresswell & Hughes (1996: Ch.11). For conditionals generally, see The Logic of Conditionals and Nute (1980b).

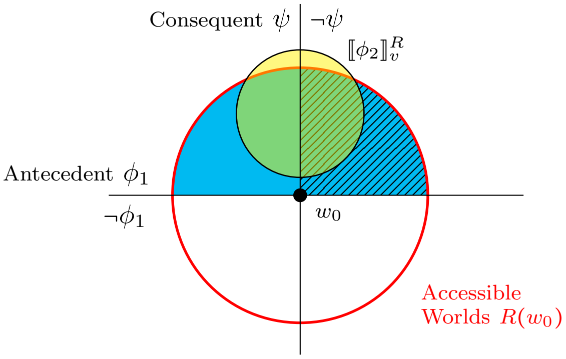

Figure 6: Strict Conditionals are Antecedent Monotonic

true, then the shaded blue region is empty, and the position of ϕ_2 reflects the fact that $[\![\phi_2]\!]_v^R \subseteq [\![\phi_1]\!]_v^R$ — recall that all worlds above the x-axis are ϕ_1-worlds. Since the shaded blue region within ϕ_2 is also empty, all ϕ_2 worlds in $R(w)$ are ψ-worlds. That is, $\phi_2 \dashv\!\!\!\!\!\!\!\;\; \psi$ is true.

Recall that Transitivity and Contraposition entail antecedent monotonicity, so it remains to show that both hold for the strict conditional. To see why Contraposition holds for the strict conditional, note again that if $\phi \dashv\!\!\!\!\!\!\!\;\; \psi$ is true in w, then all ϕ-worlds in $R(w)$ are ψ-worlds, as depicted in the left Venn diagram in Figure 7. Now suppose w is a $\neg\psi$-world in $R(w)$. As the diagram makes clear, w has to be a $\neg\phi$-world, and so $\neg\psi \dashv\!\!\!\!\!\!\!\;\; \neg\phi$ must be true in w. Similarly, if $\neg\psi \dashv\!\!\!\!\!\!\!\;\; \neg\phi$ is true in w, then all $\neg\psi$-worlds in $R(w)$ are $\neg\phi$-worlds, as depicted in the right Venn diagram in Figure 7. Now suppose w is a ϕ-world in $R(w)$. As depicted, w has to be a ψ-world, and so $\phi \dashv\!\!\!\!\!\!\!\;\; \psi$ must be true in w.

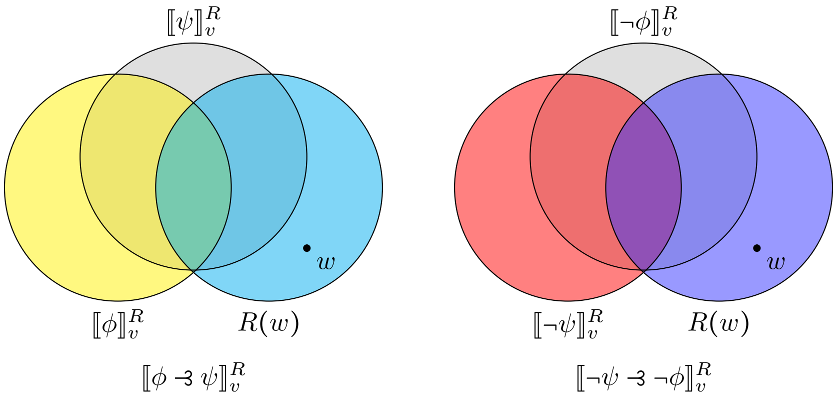
$[\![\psi]\!]_v^R$
w
$[\![\phi]\!]_v^R$
$R(w)$
$[\![\phi \dashv\!\!\!\!\!\!\!\;\; \psi]\!]_v^R$
$[\![\neg\phi]\!]_v^R$
w
$[\![\neg\psi]\!]_v^R$
$R(w)$
$[\![\neg\psi \dashv\!\!\!\!\!\!\!\;\; \neg\phi]\!]_v^R$

Figure 7: $w \in [\![\phi \dashv\!\!\!\!\!\!\!\;\; \psi]\!]_v^R \iff w \in [\![\neg\psi \dashv\!\!\!\!\!\!\!\;\; \neg\phi]\!]_v^R$ (Contraposition)

The validity of Transitivity for the strict conditional is also easy to see with a Venn diagram. The premises guarantee that all ϕ_2-worlds in $R(w)$ are ϕ_1-worlds, and that all ϕ_1-worlds in $R(w)$ are ψ-worlds. That gives one the relationships depicted in Figure 8. To show that $\phi_2 \strictif \psi$ follows, suppose that w is a ϕ_2-world in $R(w)$. As Figure 8 makes evident, w must then be a ψ-world.

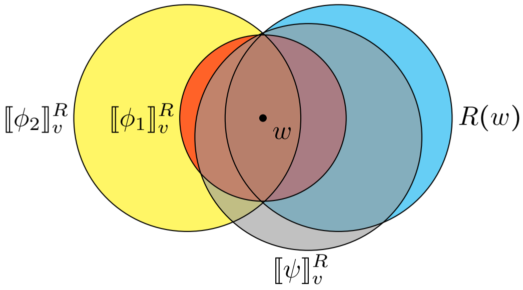

Figure 8: $w \in [\![\phi_2 \strictif \phi_1]\!]_v^R \cap [\![\phi_1 \strictif \psi]\!]_v^R \Longrightarrow w \in [\![\phi_2 \strictif \psi]\!]_v^R$ (Transitivity)

Antecendent monotonic patterns are an ineliminable part of a strict conditional logic. Examples of them often sound compelling. For example, the transitive inference (25) sounds perfectly reasonable, as does the antecedent strengthening inference (26).

(25) a. If the switch had been flipped, the light would be on.
b. If the light had been on, it would not have been dark.
c. So, if the switch had been flipped, it would not have been dark.

(26) a. If the switch had been flipped, the light would be on.
b. So, if the switch had been flipped and I had been in the room, the light would be on.

Similar examples for SDA and Contraposition as easy to find. However, counterexamples to each of the four patterns have been offered.

Counterexamples to Antecedent Strengthening were already discussed back in §1.4. Against Transitivity, Stalnaker (1968: 48) points out that (27c) does not intuitively follow from (27a) and (27b).

(27) a. If J. Edgar Hoover were today a communist, then he would be a traitor.
b. If J. Edgar Hoover had been born a Russian, then he would today be a communist.
c. If J. Edgar Hoover had been born a Russian, he would be a traitor.

Contra Contraposition, Lewis (1973b: 35) presents (28).

(28) a. If Boris had gone to the party, Olga would still have gone.
b. If Olga had not gone, Boris would still not have gone.

Suppose Boris wanted to go, but stayed away to avoid Olga. Then (28b) is false. Further suppose that Olga would have been even more excited to attend if Boris had. In that case (28a) is true. Against SDA, McKay & van Inwagen (1977: 354) offer:

(29) a. If Spain had fought for the Axis or the Allies, she would have fought for the Axis.
b. If Spain had fought for the Allies, she would have fought for the Axis.

(29b) does not intuitively follow from (29a).

These counterexamples have been widely taken to be conclusive evidence against the strict analysis (e.g. Lewis 1973b; Stalnaker 1968), since they follow from the core assumptions of that analysis. As a result, Lewis (1973b) and Stalnaker (1968) developed similarity analyses which build the non-montonicity of antecendents into the semantics of counterfactuals — see §2.3. However, there was a subsequent wave of strict analyses designed to systematically address these counterexamples. In fact, they do so by unifying two features of counterfactuals: the non-monotonic interpretation of their antecedents and their context-sensitivity.

2.2.1 Second Wave Strict Conditional Analyses

Beginning with Daniels & Freeman (1980) and Warmbrōd (1981a,b), there was a second wave of strict analyses developed explicitly to address the non-monotonic interpretation of counterfactual antecedents. Warmbrōd (1981a,b), Lowe (1983, 1990) and Lycan (2001) account for the counterexamples to antecedent monotonic patterns within a systematic theory of how counterfactuals are context-sensitive. More recently, Gillies (2007) has argued that a strict analysis along those lines is actually preferable to an account that builds the non-montonicity of counterfactual antecedents into their semantics, i.e. similarity analyses. This section will outline the basic features of these second wave strict conditional analyses.

The key idea in Warmbrōd (1981a,b) is that the accessibility sphere in the basic strict analysis should be viewed as a parameter of the context. Roughly, the idea is that $R(w)$ corresponds to background facts assumed by the participants of a discourse context. For example, if they are assuming propositions (modeled as sets of possible worlds) A, B and C then $R(w) = A \cap B \cap C$. The other key idea is that trivial strict conditionals are not pragmatically useful in conversation. If a strict conditional $\mathsf{A} \strictif \mathsf{C}$ is asserted in a context with background facts $R(w)$ and A is inconsistent with $R(w)$ — $[\![\mathsf{A}]\!]^R_v \cap R(w) = \varnothing$, then asserting $\mathsf{A} \strictif \mathsf{C}$ does not provide any information. If there are no A-worlds in $R(w)$, then, trivially, all A-worlds in $R(w)$ are C-worlds. Warmbrōd (1981a,b) proposes that conversationalists adapt a pragmatic rule of charitable interpretation to avoid trivialization:

(P) If the antecedent ϕ of a conditional is itself consistent, then $R(w) \cap [\![\phi]\!]^R_v$ should be consistent.

On this view, $R(w)$ may very well change over the course of a discourse as a result of conversationalists adhering to (P). This part of the view is central to explaining away counterexamples to antecedent monotonic validities.

Consider again the example from Goodman (1947) that appeared to be a counterexample to Antecedent Strengthening.

(30) a. If I had struck this match, it would have lit.
 b. If I had struck this match and done so in a room without oxygen, it would have lit.

Now note that if (30a) is going to come out true, the proposition that there is oxygen in the room O must be true in all worlds in the initial accessibility sphere $R_0(w)$. However, if (30b) is interpreted against $R_0(w)$, the antecedent will be inconsistent with $R_0(w)$ and so express a trivial, uninformative proposition. Warmbrōd (1981a,b) proposes that in interpreting (30b) we are forced by (P) to adopt a new, modified accessibility sphere $R_1(w)$ where O is no longer assumed. But if this is right, (30a) and (30b) don't constitute a counterexample to Antecedent Strengthening because they are interpreted against different accessibility spheres. It's like saying *All current U.S. presidents are intelligent* doesn't entail *All current U.S. presidents are unintelligent* because this sentence before Donald Trump was sworn in was true, but uttering it afterwards was false. There is an equivocation of context, or so Warmbrōd (1981a,b) contends.

Warmbrōd (1981a,b) outlines parallel explanations of the counterexamples presented to SDA, Contraposition and Transitivity. This significantly complicates the issue of whether antecedent montonicity is the key issue in understanding the semantics of counterfactuals. It appears that the non-monotonic interpretation of counterfactual antecedents can either be captured pragmatically in the way that accessibility spheres change in context (Warmbrōd 1981a,b), or it can be captured semantically as we will see from similarity analyses in §2.3. There are significant limitations to Warmbrōd's (1981a; 1981b) analysis: it does not capture nested conditionals, and does not actually predict how $R(w)$ evolves to satisfy (P). von Fintel (2001) and Gillies (2007) offer accounts that remove these limitations, and pose a challenge for traditional similarity analyses.

von Fintel (2001) and Gillies (2007) propose analyses where counterfactuals have strict truth-conditions, but they also have a dynamic meaning which effectively changes $R(w)$ non-montonically. They argue that such a theory can better explain particular phenomena. Chief among them is reverse Sobel sequences. Recall the sequence of counterfactuals (21) presented by Lewis (1973c: 419; 1973b: 10), and attributed to Howard Sobel. Reversing these sequences is not felicitous:

(31) a. If I had shirked my duty and you had too, harm would have ensued.
 $(\mathsf{I} \wedge \mathsf{U}) > \mathsf{H}$
 b. # If I had shirked my duty, no harm would have ensued.
 $\mathsf{I} > \neg\mathsf{H}$

von Fintel (2001) and Gillies (2007) observe that similarity analyses render sequences like (31) semantically consistent. Their theories predict this infelicity by providing a theory of how counterfactuals in context can change $R(w)$. Unlike von Fintel (2001), Gillies (2007) does not rely essentially on a similarity ordering over possible worlds to compute these changes to $R(w)$, and so clearly counts as a second wave strict analysis.[28]

The debate over whether counterfactuals are best given a strict or similarity analysis is very much ongoing. Moss (2012), Starr (2014) and Lewis (2017a) have proposed three different ways of explaining reverse Sobel sequences within a similarity analysis. But Willer (2015, 2017b,a) has argued on the basis of other data that a dynamic second wave strict analysis is preferable. This argument takes one into a logical comparison of strict and similarity analyses, which will be taken up in §2.4 after the similarity analysis has been presented in more detail.

2.3 Similarity Semantics

Recall the rough idea of the similarity analysis sketched in §2.1: worlds can be ordered by their similarity to the actual world, and counterfactuals say that the most similar — or least different — worlds where the antecedent is true are worlds where the consequent is also true. This idea is commonly attributed to David Lewis and Robert Stalnaker, but the actual history is a bit more nuanced. Although publication dates do not tell the full story, the approach was developed roughly contemporaneously by Stalnaker (1968), Stalnaker & Thomason (1970), Lewis (1973b), Nute (1975b) and Sprigge (1970).[29] And, there is an even earlier statement of the view:

> When we allow for the possibility of the antecedent's being true in the case of a counterfactual, we are hypothetically substituting a different world for the actual one. It has to be supposed that this hypothetical world is as much like the actual one as possible so that we will have grounds for saying that the consequent would be realized in such a world. (Todd 1964: 107)

Recall the major difference between this proposal and the basic strict analysis: the similarity analysis uses a graded notion of similarity instead of an absolute notion of accessibility. It also allows most similar worlds to vary between counterfactuals with different antecedents. These differences invalidate antecedent monotonic inference patterns. This section will introduce similarity analyses in a bit more formal detail and describe the differences between analyses within this family.

[28] See Gillies (2007: 335 fn10).

[29] The contemporaneity of Sprigge, Lewis and Stalnaker is stated in a letter from Lewis to Sprigge published in Sprigge (2006). According to Nute (1975b: 773n3), Nute (1975a) was accepted before the appearance of Lewis (1973b), but Nute (1975a) discusses the already-published Stalnaker (1968) and Stalnaker & Thomason (1970). I am indebted to a discussion of these issues by Marcus Arvan, Jessica Wilson, David Balcarras, Benj Hellie and Christopher Gauker on the blog *Philosophers' Cocoon* as part of the Campaign for Better Citation and Credit-Giving Practices in Philosophy sub-blog.

The similarity analysis has come in many varieties and formulations, including the system of spheres approach informally described in §2.1. That formulation is easiest for comparison to strict analyses. But there is a different formulation that is more intuitive and better facilitates comparison among different similarity analyses. This formulation appeals to a (set) selection function f, which takes a world w, a proposition p and returns the set of p-worlds most similar to w: $f(w,p)$.[30] $\phi > \psi$ is then said to be true when the most f-similar ϕ-worlds to w are ψ-worlds, i.e. every world in $f(w, [\![\phi]\!]^f_v)$ is in $[\![\psi]\!]^f_v$. The basics of this approach can be summed up thus.

Similarity Analysis

- $\phi > \psi$ is true at w just in case all ϕ-worlds most **similar** to w are ψ-worlds
  - Most similar according to the **selection function** f
  - f takes a proposition p and a world w and returns the p-worlds most similar to w
- $[\![\phi > \psi]\!]^f_v = \{w \mid f(w, [\![\phi]\!]^f_v) \subseteq [\![\psi]\!]^f_v\}$
- Making 'limit assumption': ϕ-worlds do not get indefinitely more and more similar to w

(Stalnaker 1968; Lewis 1973b; Nute 1975a; Pollock 1976)

As noted, this formulation makes the limit assumption: ϕ-worlds do not get indefinitely more and more similar to w. While Lewis (1973b) rejected this assumption, adopting it will serve exposition. It is discussed at length in the supplement Formal Constraints on Similarity. The logic of counterfactuals generated by a similarity analysis will depend on the constraints imposed on f. Different theorists have defended different constraints. Table 1 lists them, where $p, q \subseteq W$ and $w \in W$:

(a)	$f(w,p) \subseteq p$	**success**
(b)	$f(w,p) = \{w\}$, if $w \in p$	**strong centering**
(c)	$f(w,p) \subseteq q$ & $f(w,q) \subseteq p \implies f(w,p) = f(w,q)$	**uniformity**
(d)	$f(w,p)$ contains *at most* one world	**uniqueness**

Table 1: Candidate Constraints on Selection Functions

[30] See Lewis (1973b: §2.7) for a translation from set selection functions to using system-of-spheres formulations. Stalnaker (1968) uses a **world** selection function which by definition requires uniqueness. It also requires positing an 'absurd world' to return when p is contradictory. Nute (1975b) uses set selection functions, and Nute (1975a: 777) contends that his formulation does not require the limit assumption, and therefore contradicts Lewis (1973b: 58), who says that the system-of-spheres approach is more general than the set selection approach because the latter requires the limit assumption. This technical issue needs further investigation.

Modulo the limit assumption, Table 2 provides an overview of which analyses have adopted which constraints. Success simply enforces that $f(w,p)$ is indeed a set of p-worlds. Recall that $f(w,p)$ is supposed to be the set of most similar p-worlds to w. The other constraints correspond to certain logical validities, as detailed in the supplement Constraints on Similarity. This means that Pollock (1976) endorses the weakest logic for counterfactuals and Stalnaker (1968) the strongest. It is worth seeing how, independently of constraints (b)-(d), this semantics invalidates an antecedent monotonic pattern like Antecedent Strengthening.

Similarity Analysis	**Constraints Adopted**
Pollock (1976)	success, strong centering
Lewis (1973b), Nute (1975a)	success, strong centering, uniformity
Stalnaker (1968)	success, strong centering, uniformity, uniqueness

Table 2: Similarity Analyses, modulo Limit Assumption

Consider an instance of Antecedent Strengthening involving $\mathsf{A} > \mathsf{C}$ and $(\mathsf{A} \wedge \mathsf{B}) > \mathsf{C}$, and where the space of worlds is that given in Table 3. Now evaluate $\mathsf{A} > \mathsf{C}$ and $(\mathsf{A} \wedge \mathsf{B}) > \mathsf{C}$ in w_5 using a selection function f_1 with the following features:

World	A	B	C
w_0	1	1	1
w_1	1	1	0
w_2	1	0	1
w_3	1	0	0
w_4	0	1	1
w_5	0	1	0
w_6	0	0	1
w_7	0	0	0

Table 3: A space of worlds W, and truth-values at each world

1. $f_1(w_5, [\![\mathsf{A}]\!]_v^{f_1}) = \{w_2\}$
2. $f_1(w_5, [\![\mathsf{A} \wedge \mathsf{B}]\!]_v^{f_1}) = \{w_1\}$

Since C is true in w_2, $\mathsf{A} > \mathsf{C}$ is true in w_5 according to f_1. But, since C is false in w_1, $(\mathsf{A} \wedge \mathsf{B}) > \mathsf{C}$ is false in w_5 according to f_1. No constraints are needed here other than success. While f_1 satisfies uniqueness, the counterexample works just as well if, say, $f_1(w_5, [\![\mathsf{A}]\!]_v^{f_1}) = \{w_2, w_0\}$. Accordingly, all similarity analyses allow for the non-monotonic interpretation of counterfactual antecedents.

While Stalnaker (1968) and Lewis (1973b) remain the most popular similarity analyses, there are substantial logical issues which separate similarity analyses. These issues, and the constraints underlying them, are detailed in the supplement Formal Constraints on Similarity. Table 4 summarizes which validities go with which constraints. A few comments are in order here, though.

Constraint	**Validity**
Strong Centering	*Modus Ponens* $\phi > \psi, \phi \vDash \psi$ *Conjunction Conditionalization* $\phi \wedge \psi \vDash \phi > \psi$
Uniformity	*Substitution of Subjunctive Equivalents (SSE)* $\phi_1 > \phi_2, \phi_2 > \phi_1, \phi_1 > \psi \vDash \phi_2 > \psi$ *Limited Transitivity (LT)* $\phi_1 > \phi_2, (\phi_1 \wedge \phi_2) > \psi \vDash \phi_1 > \psi$ *Limited Antecedent Strengthening (LAS)* $\phi_1 > \phi_2, \neg(\phi_1 > \neg\psi) \vDash (\phi_1 \wedge \phi_2) > \psi$
Uniquenesss	*Conditional Excluded Middle* $\vDash (\phi > \psi) \vee (\phi > \neg\psi)$ *Conditional Negation (CN)* $\Diamond\phi \wedge \neg(\phi > \psi) \dashv\vDash \Diamond\phi \wedge \phi > \neg\psi$ *Consequent Distribution (CD)* $\phi > (\psi_1 \vee \psi_2) \vDash (\phi > \psi_1) \vee (\phi > \psi_2)$
Limit Assumption	*Infinite Consequent Entailment* If $\Gamma = \{\phi_2, \phi_3, \dots\}$, $\phi_1 > \phi_2, \phi_1 > \phi_3, \dots$ are true and $\Gamma \vDash \psi$ then $\phi_1 > \psi$

Table 4: Selection Constraints & Associated Validities

Strong centering is sufficient but not necessary for Modus Ponens, weak centering would do: $w \in f(w,p)$ if $w \in p$. LT and LAS follow from SSE, and allow similarity theorists to say why some instances of Transitivity and Antecedent Strengthening are intuitively compelling.

2.4 Comparing the Logics

The issue of whether a second wave strict analysis (§2.2.1) or a similarity analysis provides a better logic of counterfactuals is very much an open and subtle issue. As sections 2.2.1 and 2.3 detailed, both analyses have their own way of capturing the non-monotonic interpretation of antecedents. Both analyses also have their own way of capturing instances of monotonic inferences that do sound good. Perhaps this issue is destined for a stalemate.[31] But before declaring it such,

[31] Walters (2014) and Morreau (2009) try to break this stalemate, in favor of similarity analyses. But their arguments are not completely decisive. While Morreau's (2009: 447–8) counterexample to Transitivity differs from Stalnaker's (1968) (27), it comes down to whether or not it is true to assert *If it had rained, there would have been an ordinary rain shower, not a thunderstorm* in a context where a thunderstorm is a real, but unlikely, possibility. My intuitions on this aren't clear. Even if it is intuitively true, a strict theorist can say it is due to accommodating a presupposition: that we can rule out the unlikely possibility. Walters (2014) asks us to consider a context where I went to a show and my view was obstructed, and I didn't see it. Intuitively, (32a) is true. Walters (2014) argues that (32b) must be true because the consequent is true and the antecedent and consequent are independent of each

it is important to investigate two patterns that are potentially more decisive: Simplification of Disjunctive Antecedents (SDA), and a pattern not yet discussed called Import-Export.

Both SDA and Import-Export are valid in a strict analyses and invalid on standard similarity analyses. Crucially, the counterexamples to them that have been offered by similarity theorists are significantly less compelling than those offered to patterns like Antecedent Strengthening. Import-Export relates counterfactuals like (33a) and (33b).

(33) a. If Jean-Paul had danced and Simone had drummed, there would have been a groovy party.
b. If Jean-Paul had danced, then if Simone had drummed, there would have been a groovy party.

It is hard to imagine one being true without the other. The basic strict analysis agrees: it renders them equivalent.

Import-Export $(\phi_1 \wedge \phi_2) > \psi \mathrel{=\!\!|\!\!\models} \phi_1 > (\phi_2 > \psi)$

But it is not valid on a similarity analysis.[32] While Import-Export is generally regarded as a plausible principle, some have challenged it. Kaufmann (2005: 213) presents an example involving indicative conditionals which can be adapted to subjunctives. Consider a case where there is a wet match which will light if tossed in the campfire, but not if it is struck. It has not been lit. Consider now:

(34) a. If this match had been lit, it would have been lit if it had been struck.
b. If this match had been struck and it had been lit, it would have been lit.

One might then deny (34a). This match would not have lit if it had been struck, and if it had lit it would have to have been thrown into the campfire. (34b), on the other hand, seems like a straightforward logical truth. However, it is worth noting that this intuition about (34a) is very fragile. The slight variation of (34a) in (35) is easy to hear as true.

(35) If this match had been lit, then if it had been struck it (still) would have been lit.

other. Obviously, (32c) is not true, although it follows from (32a) and (32b) by Transitivity.

(32) a. If I had been an inch taller than I actually am, I would have seen the show.
b. I would not have been an inch taller than I actually am if I had seen the show.
c. If I had been an inch taller than I actually am, I would not be an inch taller than I actually am.

But, I have a hard time hearing (32b) as true, precisely because (32a)'s truth makes it hard to regard being an inch taller and me seeing the show as independent.

[32] On a similarity analysis, when the nested counterfactual $\phi_2 > \psi$ is evaluated in $\phi_1 > (\phi_2 > \psi)$, it is free to select $\neg\phi_1$-worlds. So $\phi_1 > (\phi_2 > \psi)$ will not guarantee that all most similar $\phi_1 \wedge \phi_2$-worlds are ψ-worlds.

This subtle issue may be moot, however. Starr (2014) shows that a dynamic semantic implementation of the similarity analysis can validate Import-Export, so it may not be important for settling between strict and similarity analyses.

As for the Simplification of Disjunctive Antecedents (SDA), Fine (1975), Nute (1975b), Loewer (1976) and Warmbrōd (1981a) each object to the similarity analysis predicting that this pattern is invalid. Counterexamples like (29) from McKay & van Inwagen (1977: 354) have a suspicious feature.

(29) a. If Spain had fought for the Axis or the Allies, she would have fought for the Axis.
b. # If Spain had fought for the Allies, she would have fought for the Axis.

Starr (2014: 1049) and Warmbrōd (1981a: 284) observe that (29a) seems to be another way of saying that Spain would never have fought for the Allies. While Warmbrōd (1981a: 284) uses this to pragmatically explain-away this counterexample to his strict analysis, Starr (2014: 1049) makes a further critical point: it sounds inconsistent to say (29a) after asserting that Spain could have fought for the Allies.

(36) Spain didn't fight for either the Allies or the Axis. She really could have fought for the Allies. # But, if she had fought for the Axis or the Allies, she would have fought for the Axis.

Starr (2014: 1049) argues that this makes it inconsistent for a similarity theorist to regard this as a counterexample to SDA. On a similarity analysis of the *could* claim, it follows that there are no worlds in which Spain fought for the Allies most similar to the actual world: $f(w_@, [\![\mathsf{Allies}]\!]) = \varnothing$. But if that's the case, then (29b) is vacuously true on a similarity analysis, and so a similarity theorist cannot consistently claim that this is a case where the premise is true and conclusion false. It is, however, too soon for the strict theorist to declare victory. Nute (1980a), Alonso-Ovalle (2009) and Starr (2014: 1049) each develop similarity analyses where disjunction is given a non-Boolean interpretation to validate SDA without validating the other antecedent monotonic patterns. But even this is not the end of the SDA debate.

Nute (1980b: 33) considers a similar antecedent simplification pattern involving negated conjunctions:

Simplification of Negated Conjuctive Antecedents (SNCA)

$\neg(\phi_1 \wedge \phi_2) > \neg\psi \vDash (\neg\phi_1 > \psi) \wedge (\neg\phi_2 > \psi)$

Nute (1980b: 33) presents (37) in favor of SNCA.

(37) a. If Nixon and Agnew had not both resigned, Ford would never have become President. $\neg(\mathsf{N} \wedge \mathsf{A}) > \neg\mathsf{F}$
b. If Nixon had not resigned, Ford would never have become President. $\neg\mathsf{N} > \neg\mathsf{F}$
c. If Agnew had not resigned, Ford would never have become President. $\neg\mathsf{A} > \neg\mathsf{F}$

Note that $\neg(\mathsf{N} \wedge \mathsf{A})$ and $\neg\mathsf{N} \vee \neg\mathsf{A}$ are Boolean equivalents. However, non-Boolean analyses like Nute (1980a), Alonso-Ovalle (2009) and Starr (2014: 1049) designed to capture SDA break this equivalence, and so fail to predict that SNCA is valid. Willer (2015, 2017b) develops a dynamic strict analysis which validates both SDA and SNCA. Fine (2012a,b) advocates for a departure from possible worlds semantics altogether in order to capture both SDA and SNCA. However, these accounts also face counterexamples. Fine (2012a,b) and Willer (2015, 2017b) render $(\neg\phi_1 \vee \neg\phi_2) > \psi$ and $\neg(\phi_1 \wedge \phi_2) > \psi$ equivalent, while Champollion *et al.* (2016) present a powerful counterexample to this equivalence.

Champollion *et al.* (2016) consider a light which is on when switches A and B are both up, or both down. Currently, both switches are up, and the light is on. Consider (38a) and (38b) whose antecedents are Boolean equivalents:

(38) a. If Switch A or Switch B were down, the light would be off.
$(\neg\mathsf{A} \vee \neg\mathsf{B}) > \neg\mathsf{L}$
b. If Switch A and Switch B were not both up, the light would be off.
$\neg(\mathsf{A} \wedge \mathsf{B}) > \neg\mathsf{L}$

While (38a) is intuitively true, (38b) is not.[33] This is not a counterexample to SNCA, since the premise of that pattern is false. But such a counterexample is not hard to think up.[34]

Suppose the baker's apprentice completely failed at baking our cake. It was burnt to a crisp, and the thin, lumpy frosting came out puke green. The baker planned to redecorate it to make it at least look delicious, but did not have time. We may explain our extreme dissatisfaction by asserting (39a). But the baker should not infer (39b) and assume that his redecoration plan would have worked.

(39) a. If the cake had not been burnt to a crisp and ugly, we would have been happy. $\neg(\mathsf{B} \wedge \mathsf{U}) > \mathsf{H}$
b. If the cake had not been ugly, we would have been happy.
$\neg\mathsf{U} > \mathsf{H}$

Willer (2017b: §4.2) suggests that such a counterexample trades on interpreting $\neg(\mathsf{B} \wedge \mathsf{U}) > \mathsf{H}$ as $\neg\mathsf{B} \wedge \neg\mathsf{U}) > \mathsf{H}$, and provides an independent explanation of this on the basis of how negation and conjunction interact. If this is right, then an analysis which validates SDA and SNCA without rendering $\neg(\phi_1 \wedge \phi_2) > \psi$ and $\neg\phi_1 \vee \neg\phi_2 > \psi$ equivalent is what's needed. Ciardelli *et al.* (2018) develop just such an analysis. As Ciardelli *et al.* (2018: §6.4) explain, SDA and SNCA turn out to be valid for very different reasons. Champollion *et al.* (2016); Ciardelli *et al.* (2018) also argue that the falsity of (38b) cannot be predicted on a similarity analysis. This example must be added to a long list of examples which have been presented not as counterexamples to the logic of the similarity analysis, but to what it predicts (or fails to predict) about the truth of particular

[33] Champollion *et al.* (2016) experimentally confirm that this is a robust intuition across English speakers.

[34] A counterexample to SNCA is also mentioned by Willer (2017b: §4.2), and attributed to an anonymous referee. But (39) brings out the intuition more robustly.

counterfactuals in particular contexts. This will be the topic of §2.5, where it will also be explained why the strict analysis faces similar challenges.

Where does this leave us in logical the debate between strict and similarity analyses of counterfactuals? Even Import-Export and SDA fail to clearly identify one analysis as superior. It is possible to capture SDA on either analysis. Existing similarity analyses that validate SDA, however, also invalidate SNCA (Alonso-Ovalle 2009; Starr 2014). By contrast existing strict analyses that validate SDA also validate SNCA (Willer 2015, 2017b). However, this is far from decisive. The validity of SNCA is still being investigated, and it is far from clear that it is impossible to have a similarity analysis that validates both SDA and SNCA, or a strict analysis that validates only SDA (perhaps using a non-Boolean semantics for disjunction). So even SNCA may fail to be the conclusive pattern needed to separate these analyses.

2.5 Truth-Conditions Revisited

In their own ways, Stalnaker (1968, 1984) and Lewis (1973b) are candid that the similarity analysis is not a complete analysis of counterfactuals. As should be clear from §2.3, the formal constraints they place on similarity are quite minimal and only serve to settle matters of logic. There are, in general, very many possible selection functions — and corresponding conceptions of similarity — for any given counterfactual. To explain how a given counterfactual like (40) expresses a true proposition, a similarity analysis must specify which particular conception of similarity informs it.

(40) If my computer were off, the screen would be blank.

Of course, the strict analysis is in the same position. It cannot predict the truth of (40) without specifying a particular accessibility relation. In turn, the same question arises: on what basis do ordinary speakers determine some worlds to be accessible and others not? This section will overview attempts to answer these questions, and the many counterexamples those attempts have invited. These counterexamples have been a central motivation for pursuing alternative semantic analyses, which will be covered in §3. While this section follows the focus of the literature on the similarity analysis (§2.5.1), §2.5.2 will briefly detail how parallel criticisms apply to strict analyses.

2.5.1 Truth-Conditions and Similarity

What determines which worlds are counted as most similar when evaluating a counterfactual? Stalnaker (1968) explicitly sets this issue aside, but Lewis (1973b: 92) makes a clear proposal:

Lewis' (1973b: 92) Proposal Our familiar, intuitive concept of comparative overall similarity, just applied to possible worlds, is employed in assessing counterfactuals.

Just as counterfactuals are context-dependent and vague, so is our intuitive notion of overall similarity. In comparing cost of living, New York and San Francisco may count as similar, but not in comparing topography. And yet, Lewis' (1973b: 92) Proposal has faced a barrage of counterexamples. Lewis and Stalnaker parted ways in their responses to these counterexamples, though both grant that Lewis' (1973b: 92) Proposal was not viable. Stalnaker (1984: Ch.7) proposes *the projection strategy*: similarity is determined by the way we 'project our epistemic policies onto the world'. Lewis (1979) proposes a new system of weights that amounts to a kind of curve-fitting: we must first look to which counterfactuals are intuitively true, and then find ways of weighting respects of similarity — however complex — that support the truth of counterfactuals. Since Lewis' (1973b: 92) Proposal and Lewis' (1979) system of weights are more developed, and have received extensive critical attention, they will be the focus of this section.[35] It will begin with the objections to Lewis' (1973b: 92) Proposal that motivated Lewis' (1979) system of weights, and then some objections to that approach.

Fine (1975: 452) presents the **future similarity objection** to Lewis' (1973b: 92) Proposal. (41) is plausibly a true statement about world history.

(41) If Nixon had pressed the button there would have a been a nuclear holocaust ($\mathsf{B} > \mathsf{H}$)

Suppose, optimistically, that there never will be a nuclear holocaust. Then, for every $\mathsf{B} \wedge \mathsf{H}$-world, there will be a more similar $\mathsf{B} \wedge \neg\mathsf{H}$-world, one where a small difference prevents the holocaust, such as a malfunction in the electrical detonation system. In short, a world where Nixon presses the button and a malfunction prevents a nuclear holocaust is more like our own than one where there is a nuclear holocaust that changes the face of the planet. But then Lewis' (1973b: 92) Proposal incorrectly predicts that (41) is false.

Tichý (1976: 271) offers a similar counterexample. Given (42a)-(42c), (42d) sounds false.

(42) a. Invariably, if it is raining, Jones wears his hat.
b. If it is not raining, Jones wears his hat at random.
c. Today, it is raining and so Jones is wearing his hat.
d. But, even if it had not been raining, Jones would have been wearing his hat.

Lewis' (1973b: 92) Proposal does not seem to predict the falsity of (42d). After all, Jones is wearing his hat in the actual world, so isn't a world where it's not raining and he's wearing his hat more similar to the actual one than one where it's not raining and he isn't wearing his hat?

Lewis (1979: 472) responds to these examples by proposing a ranked system of weights that give what he calls *the standard resolution of similarity*, which may be further modulated in context:

[35] The only secondary literature on Stalnaker's (1984: Ch.7) projection strategy is Pendlebury (1989) who argues that it leads to a crucially different semantics for counterfactuals.

Lewis' (1979) System of Weights

1. Avoid big, widespread, diverse violations of law. ('big miracles')
2. Maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
    - Maximize the time period over which the worlds match exactly in matters of fact
3. Avoid even small, localized, simple violations of law. ('little miracles')
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

While weight 2 gives high importance to keeping particular facts fixed up to the change required by the counterfactual, weight 4 makes clear that particular facts after that point need not be kept fixed. In the case of (42d) the fact that Jones is wearing his hat need not be kept fixed. It was a post-rain fact, so when one counterfactually supposes that it had not been raining, there is no reason to assume that Jones is still wearing his hat. Similarly, with example (41). A world where Nixon pushes the button, a small miracle occurs to short-circuit the equipment and the nuclear holocaust is prevented will count as less similar than one where there is no small miracle and a nuclear holocaust results. A small-miracle and no-holocaust world is similar to our own only in one insignificant respect (particular matters of fact) and dissimilar in one important respect (the small miracle).

It is clear, however, that Lewis' (1979) System of Weights is insufficiently general. Particular matters of fact often *are* held fixed.

(43) [You're invited to bet heads on a coin-toss. You decline. The coin comes up heads.] See, if you had bet heads you would have won! (Slote 1978: 27 fn33)

(44) If we had bought one more artichoke this morning, we would have had one for everyone at dinner tonight. (Sanford 1989: 173)

Example (43) crucially holds fixed the outcome of a highly contingent particular fact: the coin outcome. Cases of this kind are discussed extensively by Edgington (2004). Example (44) shows that a chancy outcome is not an essential feature of these cases. Noting the existence of recalcitrant cases, Lewis (1979: 472) simply says he wishes he knew why they came out differently. Additional counterexamples to the Lewis' (1979) System of Weights have been proposed by Bowie (1979), Kment (2006) and Wasserman (2006).[36] Kment (2006: 458) proposes a new similarity metric to handle this example which is sensitive to the

[36] As Francis Fairbairn (p.c.) has pointed out to me, the counterexamples proposed by Kment (2006) and Wasserman (2006) assume that Lewis' second constraint requires maximizing not just *the continuous* spatio-temporal region of exact match before a small miracle, but match over subsequent regions discontinuous with the initial one too. It is somewhat difficult to see how that interpretation of Lewis' second constraint would adequately address the original future similarity objection.

way particular facts are explained, and is integrated into a general account of metaphysical modality in Kment (2014). Ippolito (2016) proposes a new theory of how context determines similarity for counterfactuals which aims to make the correct predictions about many of the above cases.

Another response to these counterexamples has been to develop alternative semantic analyses of counterfactuals such as premise semantics (Kratzer 1989, 2012; Veltman 2005) and causal models (Schulz 2007, 2011; Briggs 2012; Kaufmann 2013). These accounts start from the observation that the counterexamples can be easily explained in a model where matters of fact depend on each other. In (42), when we counterfactually retract the fact that it rained, we don't keep the fact that the man was wearing his hat because that fact depended on it raining. Hence, (42d) is false. In (43), when we counterfactually retract that you didn't bet on heads, we keep the fact that the coin came up heads because it is independent of the fact that you didn't bet on heads. These accounts offer models of how laws, and law-like generalizations, make facts dependent on each other, and argue that once this is done, there is no work left for similarity to do in the semantics of counterfactuals. While these accounts are the focus of §3, it is worth presenting one of the additional counterexamples to the similarity analysis that has emerged from this literature.

Recall (38) from §2.4. Champollion *et al.* (2016) and Ciardelli *et al.* (2018) argue on the basis of this example that any similarity analysis will make incorrect predictions about the truth-conditions of counterfactuals. In this example a light is on either when Switch A and B are both up, or they are both down. Otherwise the light is off. Suppose both switches are up and the light is on.

(38) a. If Switch *A* or Switch *B* were down, the light would be off.
$(\neg \mathsf{A} \vee \neg \mathsf{B}) > \neg \mathsf{L}$
b. If Switch *A* and Switch *B* were not both up, the light would be off.
$\neg(\mathsf{A} \vee \mathsf{B}) > \neg \mathsf{L}$

Intuitively, (38a) is true, as are $\neg \mathsf{A} > \neg \mathsf{L}$ and $\neg \mathsf{B} > \neg \mathsf{L}$, but (38b) is false. Champollion *et al.* (2016: 321) argue that a similarity analysis cannot predict $\neg \mathsf{A} > \neg \mathsf{L}$ and $\neg \mathsf{B} > \neg \mathsf{L}$ to be true, while (38b) is false. In order for $\neg \mathsf{A} > \neg \mathsf{L}$ to be true, the particular fact that Switch *B* is up must count towards similarity. Similarly, for $\neg \mathsf{B} > \neg \mathsf{L}$ to be true, the particular fact that Switch *A* is up must count towards similarity. But then it follows that (38b) is true on a similarity analysis: the most similar worlds where *A* and *B* are not both up have to either be worlds where Switch *B* is down but Switch *A* is still up, or Switch *A* is down and Switch *B* is still up. In those worlds, the light would be off, so the similarity analysis incorrectly predicts (38b) to be true. Champollion *et al.* (2016) instead pursue a semantics in terms of causal models where counterfactually making $\neg(\mathsf{A} \wedge \mathsf{B})$ true and making $\neg \mathsf{A} \vee \neg \mathsf{B}$ true come apart.

2.5.2 Truth-Conditions and the Strict Analysis

Do strict analyses avoid the troubles faced by similarity analyses when it comes to truth-conditions? This question is difficult to answer, and has not been explicitly discussed in the literature. Other than the theory of Warmbrōd (1981b,a), strict theorists have not made proposals for the accessibility relation analogous to Lewis' (1973b: 92) Proposal for similarity. And, Warmbrōd's proposal about the pragmatics of the accessibility relation is this:

Warmbrōd's (1981a: 280) Proposal

> In the normal case of interpreting a conditional with a nonabsurd antecedent p, the worlds accessible from w will be those that are as similar to w as the most similar p-worlds.

All subsequent second wave strict analyses have ended up in similar territory. The dynamic analyses developed by von Fintel (2001), Gillies (2007) and Willer (2015, 2017b,a) assign strict truth-conditions to counterfactuals, but have them induce changes in an evolving space of possible worlds. These changes must render the antecedent consistent with an evolving body of discourse. While von Fintel (2001) and Willer (2017a) explicitly appeal to a similarity ordering for this purpose, Gillies (2007) and Willer (2017b) do not. Nevertheless, the formal structures used by Gillies (2007) and Willer (2017b) for this purpose give rise to the same question: which facts stay and which facts go when rendering the counterfactual antecedent consistent? Accordingly, at present, it does not appear that the strict analysis avoids the kinds of concerns raised for the similarity analysis in §2.5.1.

2.6 Philosophical Objections

Recall Goodman's Problem from §1.4: the truth-conditions of counterfactuals intuitively depend on background facts and laws, but it is difficult to specify these facts and laws in a way that does not itself appeal to counterfactuals. Strict and similarity analyses make progress on the logic of conditionals without directly confronting this problem. But the discussion of §§2.5 makes salient a related problem. Lewis' (1979) System of Weights amounts to reverse-engineering a similarity relation to fit the intuitive truth-conditions of counterfactuals. While Lewis' (1979) approach avoids characterizing laws and facts in counterfactual terms, Bowie (1979: 496-7) argues that it does not explain why certain counterfactuals are true without appealing to counterfactuals. Suppose one asks why certain counterfactuals are true and the similarity theorist replies with Lewis' (1979) recipe for similarity. If one asks why those facts about similarity make counterfactuals true, the similarity theorist cannot reply that they are basic self-evident truths about the similarity of worlds. Instead, they must say that those similarity facts make those counterfactuals true. Bowie's (1979: 496-7) criticism is that this is at best uninformative, and at worst circular.

A related concern is voiced by Horwich (1987: 172) who asks "why we should have evolved such a baroque notion of counterfactual dependence", namely that captured by Lewis' (1979) System of Weights. The concern has two components: why would humans find it useful, and why would human psychology ground counterfactuals in this concept of similarity rather than our ready-at-hand intuitive concept of overall similarity? These questions are given more weight given the centrality of counterfactuals to human rationality and scientific explanation outlined in §1. Psychological theories of counterfactual reasoning and representation have found tools other than similarity more fruitful (§1.2). Similarly, work on scientific explanation has not assigned any central role for similarity (1.3), and as Hájek (2014: 250) puts it: "Science has no truck with a notion of similarity; nor does Lewis' (1979) ordering of what matters to similarity have a basis in science."

Morreau (2010) has recently argued on formal grounds that similarity is poorly suited to the task assigned to it by the similarity analysis. The similarity analysis, especially as elaborated by Lewis (1979), tries to weigh some similarities between worlds against their differences to arrive at a notion of overall comparative similarity between those worlds. Morreau (2010: 471) argues that: "[w]e cannot add up similarities or weigh them against differences. Nor can we combine them in any other way... No useful comparisons of overall similarity result." Morreau (2010: §4) articulates this argument formally via a reinterpretation of Arrow's Theorem in social choice theory — see entry Arrow's Theorem. Arrow's Theorem shows that it is not possible to aggregate individuals' preferences regarding some alternative outcomes into a coherent 'collective preference' ordering over those outcomes, given minimal assumptions about their rationality and autonomy. As summarized in §6.3 of Arrow's Theorem, Morreau (2010) argues that the same applies to aggregating respects of similarity and difference: there is no way to add them up into a coherent notion of overall similarity.

2.7 Summary

Strict and similarity analyses of counterfactuals showed that it was possible to address the semantic puzzles described in §1.4 with formally explicit logical models. This dispelled widespread skepticism of counterfactuals and established a major area of interdisciplinary research. Strict analyses have been revealed to provide a stronger, more classical, logic, but must be integrated with a pragmatic explanation of how counterfactual antecedents are interpreted non-monotonically. Similarity analyses provide a much weaker, more non-classical, logic, but capture the non-monotonic interpretation of counterfactual antecedents within their core semantic model. It is now a highly subtle and intensely debated question which analysis provides a better logic for counterfactuals, and which version of each kind of analysis is best. This intense scrutiny and development has also generated a wave of criticism focused on their treatment of truth-conditions, Goodman's Problem and integration with thinking about counterfactuals in psychology and the philosophy of science (§§2.5, 2.6). None of these criticisms are absolutely conclusive, and these two analyses, particularly the similarity analysis, remain standard in philosophy and linguistics. However, the criticisms are serious enough to merit exploring alternative analyses. These alternative accounts take inspiration from a particular diagnosis of the counterexamples discussed in §2.5: facts depend on each other, so counterfactually assuming p involves not just giving up *not-p*, but any facts which depended on *not-p*. The next section will examine analyses of this kind.

3 Semantic Theories of Counterfactual Dependence

Similarity and strict analyses nowhere refer to facts, or propositions, depending on each other. Indeed, Lewis (1979) was primarily concerned with explaining which true counterfactuals, given a similarity analysis, manifest a relation of counterfactual dependence. Other analyses have instead started with the idea that facts depend on each other, and then explain how these relations of dependence make counterfactuals true. As will become clear, none of these analyses endorse the naive idea that $\mathsf{A} > \mathsf{B}$ is true only when B counterfactually depends on A. The dependence can be more complex, indirect, or B could just be true and independent of A. Theories in this family differ crucially in how they model counterfactual dependence. In premise semantics (§3.1) dependence is modeled in terms of how facts, which are modeled as parts of worlds, are distributed across a space of worlds that has been constrained by laws, or law-like generalizations. In probabilistic semantics (§3.2), this dependence is modeled as some form of conditional probability. In Bayesian networks, structural equations and causal models (§3.3), it is modeled in terms of the Bayesian networks discussed at the beginning of §1.2.3. Because theories of these three kinds are very much still in development and often involve even more sophisticated formal models than those covered in §2, this section will have to be more cursory than §2 to ensure breadth and accessibility.

3.1 Premise Semantics

Veltman (1976) and Kratzer (1981b) approached counterfactuals from a perspective closer to Goodman (1947): counterfactuals involve explicitly adjusting a body of premises, facts or propositions to be consistent with the counterfactual's antecedent, and checking to see if the consequent follows from the revised premise set — in a sense of 'follow' to be articulated carefully. Since facts or premises hang together, changing one requires changing others that depend on it. The function of counterfactuals is to allow us to probe these connections between facts. While Lewis (1981) proved that the Kratzer (1981b) analysis was a special case of similarity semantics, subsequent refinements of premise semantics in Kratzer (1989, 1990, 2002, 2012) and Veltman (2005) evidenced important differences. Kratzer (1989: 626) nicely captures the key difference:

> [I]t is not that the similarity theory says anything false about [particular] examples... It just doesn't say enough. It stays vague where our intuitions are relatively sharp. I think we should aim for a theory of counterfactuals that is able to make more concrete predictions with respect to particular examples.

From a logical point of view, premise semantics and similarity semantics do not diverge. They diverge in the concrete predictions made about the truth-conditions of counterfactuals in particular contexts without adding additional constraints to the theory like Lewis' (1979) System of Weights.

How does premise semantics aim to improve on the predictions of similarity semantics? It re-divides the labor between context and the semantics of counterfactuals to more accurately capture the intuitive truth-conditions of counterfactuals, and intuitive characterizations of how context influences counterfactuals. In premise semantics, context provides facts and law-like relations among them, and the counterfactual semantics exploits this information. By contrast, the similarity analysis assumes that context somehow makes a similarity relation salient, and has to make further stipulations like Lewis' (1979) System of Weights about how facts and laws enter into the truth-conditions of counterfactuals in particular contexts. This can be illustrated by considering how Tichý's (1976) example (42) is analyzed in premise semantics. This illustration will use the Veltman (2005) analysis because it is simpler than Kratzer (1989, 2012) — that is not to say it is preferable. The added complexity in Kratzer (1989, 2012) provides more flexibility and a broader empirical range including quantification and modal expressions other than *would*-counterfactuals.

Recall Tichý's (1976) example, with the intuitively false counterfactual (42d):

(42) a. Invariably, if it is raining, Jones wears his hat.
$\Box(\mathsf{R} \supset \mathsf{W})$
b. If it is not raining, Jones wears his hat at random.
c. Today, it is raining and so Jones is wearing his hat.
$\mathsf{R} \wedge \mathsf{W}$
d. But, even if it had not been raining, Jones would have been wearing his hat.
$\neg\mathsf{R} > \mathsf{W}$

Veltman (2005) models how the sentences leading up to the counterfactual (42d) determine the facts and laws relevant to its interpretation. The law-like generalization in (42a) is treated as a strict conditional which places a hard constraint on the space of worlds relevant to evaluating the counterfactual.[37] The particular facts introduced by (42c) provide a soft constraint on the worlds relevant to interpreting the counterfactual. Figure 9 illustrates this model of the context and its evolution, including a third atomic sentence H for reasons that will become clear shortly. On this model a context provides a set of worlds compatible with the facts, in C_2 $\mathit{Facts}_{C_2} = \{w_6, w_7\}$, and the set of worlds compatible with the laws, in C_2 $\mathit{Universe}_{C_2} = \{w_0, w_1, w_2, w_3, w_6, w_7\}$. This model of context is one essential component of the analysis, but so too is the way Veltman (2005) models worlds, situations and dependencies between facts. These further components allow Veltman (2005) to offer a procedure for 'retracting' the fact that R holds from a world.

[37] Veltman (2005) does not include (42b). Presumably it just ensures that there are both $\mathsf{R} \wedge \mathsf{W}$ and $\mathsf{R} \wedge \neg\mathsf{W}$ worlds.

C_0	R	W	H
$\boldsymbol{w_0}$	**0**	**0**	**0**
$\boldsymbol{w_1}$	**0**	**0**	**1**
$\boldsymbol{w_2}$	**0**	**1**	**0**
$\boldsymbol{w_3}$	**0**	**1**	**1**
$\boldsymbol{w_4}$	**1**	**0**	**0**
$\boldsymbol{w_5}$	**1**	**0**	**1**
$\boldsymbol{w_6}$	**1**	**1**	**0**
$\boldsymbol{w_7}$	**1**	**1**	**1**

$\xrightarrow{\Box(\mathsf{R} \supset \mathsf{W})}$

C_1	R	W	H
$\boldsymbol{w_0}$	**0**	**0**	**0**
$\boldsymbol{w_1}$	**0**	**0**	**1**
$\boldsymbol{w_2}$	**0**	**1**	**0**
$\boldsymbol{w_3}$	**0**	**1**	**1**
~~w_4~~	~~1~~	~~0~~	~~0~~
~~w_5~~	~~1~~	~~0~~	~~1~~
$\boldsymbol{w_6}$	**1**	**1**	**0**
$\boldsymbol{w_7}$	**1**	**1**	**1**

$\xrightarrow{\mathsf{R} \wedge \mathsf{W}}$

C_2	R	W	H
w_0	0	0	0
w_1	0	0	1
w_2	0	1	0
w_3	0	1	1
~~w_4~~	~~1~~	~~0~~	~~0~~
~~w_5~~	~~1~~	~~0~~	~~1~~
$\boldsymbol{w_6}$	**1**	**1**	**0**
$\boldsymbol{w_7}$	**1**	**1**	**1**

Figure 9: Context for (42), Facts in Bold, Laws Crossing out Worlds

Veltman's (2005) analysis of counterfactuals identifies possible worlds with atomic valuations (functions from atomic sentences to truth-values) like those depicted in Figure 9. So $w_6 = \{\langle \mathsf{R}, 1\rangle, \langle \mathsf{W}, 1\rangle, \langle \mathsf{H}, 0\rangle\}$. This makes it possible to offer a simple model of *situations*, which are parts of worlds: any subset of a world.[38] It is now easy to think about one fact (sentence having a truth-value) as determining another fact (sentence having a truth value). In context C_3, R being 1 determines that W will be 1. Once you know that R is assigned to 1, you know that W is too. Veltman's (2005) proposal is that speakers evaluate a counterfactual by retracting the fact that the antecedent is false from the worlds in the context, which gives you some situations, and then consider all those worlds that contain those situations, are compatible with the laws, and make the antecedent true. If the consequent is true in all of those worlds, then we can say that the counterfactual is true in (or supported by) the context. So, to evaluate $\neg\mathsf{R} > \mathsf{W}$, one first retracts the fact that R is true, i.e. that R is assigned to 1, then one finds all the worlds consistent with the laws that contain those situations and assign R to 0. If all of those worlds are also W worlds, then the counterfactual is true in (or supported by) the context. For Veltman (2005), the characterization of this retraction process relies essentially on the idea of facts determining other facts.

According to Veltman (2005), when you are 'retracting' a fact from the facts in the context, you begin by considering each $w \in \mathit{Facts}_C$ and find the smallest situations in w which contain only undetermined facts — he calls such a situation a *basis* for w. This is a minimal situation which, given the laws constraining $\mathit{Universe}_C$, determines all the other facts about that world. For example, w_6 has only one basis, namely $s_0 = \{\langle \mathsf{R}, 1\rangle, \langle \mathsf{H}, 0\rangle\}$, and w_7 has only one basis, namely $s_1 = \{\langle \mathsf{R}, 1\rangle, \langle \mathsf{H}, 1\rangle\}$. Once you have the bases for a world, you can retract a fact by finding the smallest change to the basis that no longer forces that fact to be true. So retracting the fact that R is true from s_0 produces $s_0' = \{\langle \mathsf{H}, 0\rangle\}$, and retracting it from s_1 produces $s_1' = \{\langle \mathsf{H}, 1\rangle\}$. The set consisting of these two situations is the *premise set*.

[38] Situations are an essential part of both Veltman (1985) and Kratzer (1989, 2012), although Kratzer (1989, 2012) offers a much more intricate theory of situations. See entry Situations in Natural Language Semantics.

$C_{(43)}$	B	H	W
w_0	0	0	0
~~w_1~~	~~0~~	~~0~~	~~1~~
$\boldsymbol{w_2}$	**0**	**1**	**0**
~~w_3~~	~~0~~	~~1~~	~~1~~
w_4	1	0	0
~~w_5~~	~~1~~	~~0~~	~~1~~
~~w_6~~	~~1~~	~~1~~	~~0~~
w_7	1	1	1

Figure 10: Context for (43)

To evaluate $\neg\mathsf{R} > \mathsf{W}$, one finds the set of worlds from $\mathit{Universe}_{C_3}$ that contains some member of the premise set s_0' or s_1': $\{w_0, w_1, w_2, w_3\}$ — these are the worlds consistent with the premise set and the laws. Are all of the $\neg$R-worlds in $\{w_0, w_1, w_2, w_3\}$ also W-worlds? No, w_2 and w_3 are not. Thus, $\neg\mathsf{R} > \mathsf{W}$ is not true in (or supported by) the context C_3. This was the intuitively correct prediction about example (42). Of course, the similarity analysis supplemented with Lewis' (1979) System of Weights also makes this prediction. But consider again example (43), which is not predicted:

(43) [You're invited to bet heads on a coin-toss. You decline. The coin comes up heads.] See, if you had bet heads you would have won! (Slote 1978: 27 fn33)

This example relies seamlessly on three pieces of background knowledge about how betting works:

1. If you don't bet, you don't win: $\Box(\neg\mathsf{B} \supset \neg\mathsf{W})$
2. If you bet and it comes up heads, you win: $\Box((\mathsf{B} \wedge \mathsf{H}) \supset \mathsf{W})$
3. If you bet and it doesn't come up heads, you don't win: $\Box((\mathsf{B} \wedge \neg\mathsf{H}) \supset \neg\mathsf{W})$

And it specifies facts: $\neg\mathsf{B} \wedge \mathsf{H}$. The resulting context is detailed in Figure 10: Now, consider the counterfactual $\mathsf{B} > \mathsf{W}$. The first step is to retract the fact that B is false from each world in $\mathit{Facts}_{C_{(43)}}$. That's just w_2. This world has two bases — minimal situations consisting of undetermined facts — $s_0 = \{\langle \mathsf{B}, 0\rangle, \langle \mathsf{H}, 1\rangle\}$ and $s_1 = \{\langle \mathsf{H}, 1\rangle, \langle \mathsf{W}, 0\rangle\}$.[39] The next step is to retract the fact that B is false from both bases. For s_0 this yields $s_0' = \{\langle \mathsf{H}, 1\rangle\}$ and for s_1 this also yields s_0' — since the fact that you didn't win together with the fact that the coin came up heads, forces it to be false that you bet. Given this situation, the premise set consists of the two worlds in $\mathit{Universe}_{(43)}$ that contain s_0': $\{w_2, w_7\}$. Now, are all of the B-worlds in this set also W-worlds? Yes, w_7 is the only B-world, and

[39] The fact that you didn't bet determines the fact that you didn't win, so $\langle \mathsf{B}, 0\rangle$ and $\langle \mathsf{W}, 0\rangle$ cannot both be in a basis for w_2. But neither fact determines whether the coin came up heads, and whether the coin came up heads does not determine whether you bet, or whether you won.

C_2	A	B	L
w_0	0	0	0
~~w_1~~	~~0~~	~~0~~	~~1~~
w_2	0	1	0
~~w_3~~	~~0~~	~~1~~	~~1~~
$\boldsymbol{w_4}$	**1**	**0**	**0**
~~w_5~~	~~1~~	~~0~~	~~1~~
~~w_6~~	~~1~~	~~1~~	~~0~~
w_7	1	1	1

Figure 11: Context for (45d)

it is also a W-world. So Veltman (2005) correctly predicts that (43) is true in (supported by) its natural context.

It should now be more clear how premise semantics delivers on its promise to be more predictive than similarity semantics when it comes to counterfactuals in context, and affords a more natural characterization of how a context informs the interpretation of counterfactuals. This analysis was crucially based on the idea that some facts determine other facts, and that the process of retracting a fact is constrained by these relations. However, even premise semantics has encountered counterexamples.

Schulz (2007: 101) poses the following counterexample to Veltman (2005).

(45) a. If both Switch A and Switch B are up, the light is on.
$\Box((\mathsf{A} \wedge \mathsf{B}) \supset \mathsf{L})$
b. If either Switch A or Switch B is down, the light is off.
$\Box((\neg\mathsf{A} \vee \neg\mathsf{B}) \supset \neg\mathsf{L})$
c. Switch A is up, Switch B is down, and the light is off.
$\mathsf{A} \wedge \neg\mathsf{B} \wedge \neg\mathsf{L}$
d. If Switch B had been up, the light would have been on.
$\mathsf{B} > \mathsf{L}$

Intuitively, (45d) is true in the context. Figure 11 details the context predicted for it by Veltman (2005). There are two bases for w_4: $s_0 = \{\langle \mathsf{A}, 1\rangle, \langle \mathsf{L}, 0\rangle\}$ — the fact that Switch A is up and the light is off determines that Switch B is down — and $s_1 = \{\langle \mathsf{A}, 1\rangle, \langle \mathsf{B}, 0\rangle\}$ — the fact that Switch A is up and the fact that B is down determines that the light is off. (No smaller situation would determine the facts of w_4.) Retracting B's falsity from s_0 leads to trouble. s_0 forces B to be false, but there are two ways of changing this. First, one can remove the fact that the light is on, yielding $s_0' = \{\langle \mathsf{A}, 1\rangle\}$. Second, one can eliminate the fact that Switch A is up, yielding $s_0'' = \{\langle \mathsf{L}, 0\rangle\}$. Because of s_0'', the premise set will contain w_2, meaning it allows that in retracting the fact that Switch B is down one can give up the fact that Switch A is up. But then there is a B-world where L is false, and $\mathsf{B} > \mathsf{L}$ is incorrectly predicted to be false.

Intuitively, the analysis went wrong in allowing the removal of the fact that Switch A is up when retracting the fact that Switch B is down. Schulz

(2007: §5.5) provides a more sophisticated version of this diagnosis: although the fact that Switch A is up and the fact that the light is off together determine that Switch B is down, only the fact that the light is off depends on the fact that Switch B is down. If one could articulate this intuitive concept of dependence, and instead only retract facts that depend on the fact you are retracting (in this case the fact that B is down), then the error could be avoided. It is unclear how to implement this kind of dependence in Veltman's (2005) framework. Schulz (2007: §5.5) goes on show that structural equations and causal models provide the necessary concept of dependence — for more on this approach see §3.3 below. After all, it seems plausible that the light being off causally depends on Switch B being down, but Switch A being up does not causally depend on Switch B being down. It remains to be seen whether the more powerful framework developed by Kratzer (1989, 2012) can predict (45).

3.2 Conditional Probability Analyses

While premise semantics has been prominent among linguists, probabilistic theories have been very prominent among philosophers thinking about knowledge and scientific explanation.[40] Adams (1965, 1975) made a seminal proposal in this literature:

Adams' Thesis The assertability of q *if* p is proportional to $P(q \mid p)$, where P is a probability function representing the agent's subjective credences — see Definition 1.

However, Adams (1970) was also aware that indicative/subjunctive pairs like (3)/(4) differ in their assertability. To explain this, he proposed the *prior probability* analysis of counterfactuals Adams (1976):

Adams' Prior Probability Analysis The assertability of $\phi > \psi$ is proportional to $P_0(\psi \mid \phi)$, where P_0 is the agent's credence prior to learning that ϕ was false.

It would seem that this analysis accurately predicts our intuitions in (45) about $\mathsf{B} > \mathsf{L}$. Let P_0 be an agent's credence before learning that Switch B is down. (45a) requires that $P_0(\mathsf{L} \mid \mathsf{A} \wedge \mathsf{B})$ is (or is close to) 1, (45b) requires that $P_0(\neg\mathsf{L} \mid \neg\mathsf{A} \vee \neg\mathsf{B})$ is (or is close to) 1. The agent also learns that Switch A is up, so $P_0(\mathsf{A})$ is (or is close to) 1. All of this together seems to guarantee that $P_0(\mathsf{B} \mid \mathsf{L})$ is also very high. However, this is due to an inessential artifact of the example: the agent learned that Switch B was down *after* learning that Switch A is up. This detail does not matter to the intuition. As was seen with example (43), we often hold fixed facts that happen after the antecedent turns out false. Indeed, Adams' Prior Probability Analysis makes the incorrect prediction that (43) is unassertible in its natural context.

This problem for Adams' Prior Probability Analysis is addressed in Edgington (2003, 2004: 21) who amends the analysis: P_0 may also reflect any facts the

[40] See the entry Logic and Probability for a detailed discussion of probabilistic tools.

agent learns after they learn that the antecedent is false, provides that those facts are causally independent of the antecedent. This parallels the idea pursued by Schulz (2007: Ch.5) to integrate causal dependence into the analysis of counterfactuals. This idea was also pursued in a probabilistic framework by Kvart (1986, 1992). Kvart (1986, 1992), however, does not propose a prior probability analysis and does not regard the probabilities as subjective credences: they are instead objective probabilities (propensity or objective chance). Skyrms (1981) also proposes a propensity account, but pursues a prior propensity account analogous to the subjective one proposed by Adams (1976).

Objective probability analyses have been popular among philosophers trying to capture the way that counterfactuals feature in physical explanations, and why they are so useful to agents like us in worlds like ours. Loewer (2007) is a good example of such an account, who grounds the truth of certain counterfactuals regarding our decisions like (46) in statistical mechanical probabilities.

(46) If I were to decide to bet on the coin's landing heads, then the chance I would win is 0.5

Loewer (2007) proposes that (46) is true just in case (where P_{SM} is the statistical mechanical probability distribution and $M(t)$ is a description of the macro-state of the universe at t):

(47) $P_{SM}(W \mid M(t) \wedge D_1) = 0.5$

Loewer (2007) acknowledges that this analysis is limited to counterfactuals like (46). He argues that it can address the philosophical objections to the similarity analysis discussed in §2.6, namely why counterfactuals are useful in scientific explanations, and for agents like us in a world like our own.

Conditional probability analyses do not proceed by assigning truth-conditions to (all) counterfactuals. They instead associate them with certain conditional probabilities.[41] This makes it difficult to integrate the theory into a comprehensive compositional semantics and logic for a natural language. Kaufmann (2005, 2008) makes important advances here, but it remains an open issue for conditional probability analyses. Leitgeb (2012a,b) thoroughly develops a new conditional probability analysis which regards $\phi > \psi$ as true when the relevant conditional probability is sufficiently high.[42] But conditional probability analyses have other limitations. Without further development, these analyses are limited in their ability to explain how humans judge particular counterfactuals to be true. There is a large literature in psychology, beginning with Kahneman *et al.* (1982), showing that human reasoning diverges in predictable way from precise probabilistic reasoning. Even if these performance differences didn't turn up in counterfactuals and conditional probabilities, there is an implementation

[41] Loewer (2007) finesses this issue but only by considering only counterfactuals that have an explicit probability value in the consequent.

[42] While Hawthorne (2005) has criticized probabilistic approaches for not validating Agglomeration — $\phi > \psi_1, \phi > \psi_2$ therefore $\phi > (\psi_1 \wedge \psi_2)$ — this pattern is validated by theories like Adams (1976) and Leitgeb (2012a,b).

issue. As discussed in §1.2.3, directly implementing probabilistic knowledge makes unreasonable demands on memory. Bayesian Networks are one proposed solution to this implementation issue. They are also used in the analysis of causal dependence (§1.3), which conditional probability analyses must appeal to anyway. Since Bayesian Networks can also be used to directly formulate a semantics of counterfactuals, they provide an worthwhile alternative to conditional probability analyses despite proceeding from similar assumptions.

3.3 Bayesian Networks, Structural Equations and Causal Models

Recall from §1.2.3 the basic idea of a Bayesian Network: rather than storing probability values for all possible combinations of some set of variables, a Bayesian Network represents only the conditional probabilities of variables whose values depend on each other. This can be illustrated for (45).

(45) a. If both Switch A and Switch B are up, the light is on.
b. If either Switch A or Switch B is down, the light is off.
c. Switch A is up, Switch B is down, and the light is off.
d. If Switch B had been up, the light would have been on.

Sentences (45a)-(45c) can be encoded by the Bayesian Network and structural equations in Figure 12. Recall that $L \coloneqq A \wedge B$ means that the value of L equals the value of $A \wedge B$, but also asymmetrically depends on the value of $A \wedge B$: the value of $A \wedge B$ determines the value of L, and not vice-versa. How, given the network in Figure 12, does one evaluate the counterfactual $\mathsf{B} > \mathsf{N}$? Several different answers have been given to this question.

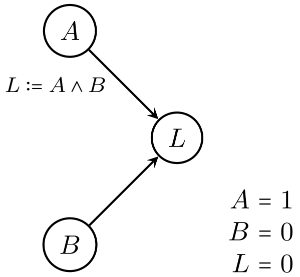

Figure 12: Bayesian Network and Structural Equations for (45)

Pearl (1995, 2000, 2009, 2013: Ch.7) proposes:

Interventionism Evaluate $\mathsf{B} > \mathsf{L}$ relative to a Bayesian Network by removing any incoming arrows to B, setting its value to 1, and projecting this change forward through the remaining network. If L is 1 in the resulting network, $\mathsf{B} > \mathsf{L}$ is true; otherwise it's false.

On this approach, one simply deletes the assignment $B = 0$, replaces it with $B = 1$, and solves for L using the equation $L \coloneqq A \wedge B$. Since the deletion of $B = 0$ does not effect the assignment $A = 1$, it follows that $L = 1$ and that the counterfactual is true. This simple recipe yields the right result. Pearl nicely sums up the difference between this kind of analysis and a similarity analysis:

> In contrast with Lewis's theory, counterfactuals are not based on an abstract notion of similarity among hypothetical worlds; instead, they rest directly on the mechanisms (or 'laws,' to be fancy) that produce those worlds and on the invariant properties of those mechanisms. Lewis's elusive 'miracles' are replaced by principled [interventions] which represent the minimal change (to a model) necessary for establishing the antecedent... Thus, similarities and priorities — if they are ever needed — may be read into the [interventions] as an afterthought... but they are not basic to the analysis. (Pearl 2009: 239-40)

As interventionism is stated above, it does not apply to conditionals with logically complex antecedents or consequents. This limitation is addressed by Briggs (2012), who also axiomatizes and compares the resultant logic to Lewis (1973b) and Stalnaker (1968) — significantly extending the analysis and results in Pearl (2009: Ch.7). Integrations of causal models with premise semantics (Schulz 2007, 2011; Kaufmann 2013; Santorio 2014; Champollion *et al.* 2016; Ciardelli *et al.* 2018) provide another way of incorporating an interventionist analysis into a fully compositional semantics. However, interventionism does face other limitations.

Hiddleston (2005) presents the following example.

(48) a. If the cannon is lit, there is a simultaneous flash and bang.
b. The cannon was not lit, there was no flash, and no bang.
c. But, if there had been a flash, there would have been a bang.

(48c) is intuitively true in this context. The network for (48) is given in Figure 13. Hiddleston (2005) observes that interventionism does not predict $\mathsf{F} > \mathsf{B}$ to be true. It tells one to delete the arrow going in to F, set its value to 1 and project

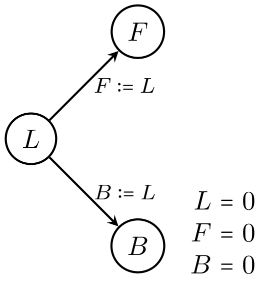

Figure 13: Bayesian Network and Structural Equations for (48)

the consequences of doing so. However, none of the other values depend on F so they keep their actual values: $L = 0$ and $B = 0$. Accordingly, $\mathsf{F} > \mathsf{B}$ is false, contrary to intuition. Further, because the intervention on F has destroyed its connection to B, it's not even possible to tweak interventionism to allow values to flow backwards (to the left) through the network.[43]

Hiddleston's (2005) counterexample highlights the possibility of another kind of counterexample featuring embedded conditionals. Consider again the network in Figure 12. The following counterfactual seems true (Starr 2012: 13).

(49) If the light had been on, then if you had flipped Switch A down, the light would be off.

And, considering a simple match, Fisher (2017a: §1) observes that (50b) is intuitively false.

(50) a. A match is struck but does not light.
b. If the match had lit, then (even) if it had not been struck, it would have lit.

In both cases, interventionism is destined to make the wrong prediction. With (49), the intervention in the first antecedent removes the connection between Switch A and the light, so when the antecedent of the consequent is made true by intervention, it does not result in L's value becoming 0. And so the whole counterfactual comes out false. Similarly with (50b), when the first antecedent is made true by intervention, it stays true even after the second antecedent is evaluated. Hence the whole conditional is predicted to be true. Fisher (2017b) also observes that interventionism also has no way of treating counterlegal counterfactuals like *if Switch A had alone controlled the light, the light would be on.*

These counterexamples to interventionism have stimulated alternative accounts like Hiddleston's (2005) *minimal network analysis* and further developments of that analysis (Rips 2010; Rips & Edwards 2013; Fisher 2017a). Instead of modifying an existing network to make the antecedent true, this analysis considers alternate networks where only the parent nodes of the antecedent which directly influence it are changed to make the antecedent come true. However, Pearl's (2009) interventionist analysis has also been incorporated into the *extended structural models* analysis (Lucas & Kemp 2015). This analysis aims to capture interventions as a special case of a more general proposal about how antecedents are made true. One important aspect of this proposal is that interventions often involve inserting a hidden node that amounts to an unknown cause of the antecedent. The analysis of Snider & Bjorndahl (2015) pursues a third idea: counterfactuals are not interpreted by manipulating a background network, but instead serve to constrain the class of possible networks compatible with the information shared in a conversation, as in Stalnaker's (1978) theory of assertion.[44] Among these relations can be cause-to-effect networks as in (45d),

[43] Although this kind of reasoning is sometimes referred to as 'backtracking', the conditional (48c) is not a backtracker in the sense of Lewis (1979): the consequent event does not occur before the antecedent event — they are simultaneous.

[44] For more on this see §5.2 of the entry Assertion.

but also networks that involve the antecedent and consequent having a common cause, as in (48c). As should be clear, this is a rapidly developing area of research where it is not possible to identify one analysis as standard or representative. It does bear emphasizing that this literature is driven not only by precise formal models, but also by experimental data which is brought to bear on the predictions of these analyses.

A few final philosophical remarks are in order about the kinds of analyses discussed here. If one follows Woodward (2002) and Hitchcock (2001) in their interpretation of these networks, a structural equation should be viewed as a primitive counterfactual. It follows that this is a non-reductive analysis of counterfactual dependence: it only explains how the truth of arbitrarily complex counterfactual sentences are grounded in basic relations of counterfactual dependence. However, note in the earlier quotation above from Pearl (2009: 239-40) that he interprets structural equations as basic mechanisms or laws, and so arguably counts as an analysis of counterfactuals in terms of laws. These amount to two very different philosophical positions that interact with the philosophical debates surveyed in §1.3.

It is also worth noting that while many working in this framework apply these networks to causal relations, there is no reason to assume that the analysis would not apply to other kinds of dependence relations. For example, constitutional dependence is at the heart of counterfactuals like:

(51) If Socrates hadn't existed, the set consisting of Socrates wouldn't have existed.

From a Bayesian Network approach to mental representation (§1.2.3), this makes perfect sense: the networks encode probabilistic dependence which can come from causal or constitutional facts.

Finally, it is worth highlighting that the philosophical objections directed at the similarity analysis in §2.6 are addressed, at least to some degree, by structural equation analyses. Because the central constructs of this analysis — structural equations and Bayesian Networks — are also employed in models of mental representation, causation and scientific explanation, it grounds counterfactuals in a construct already taken to explain how creatures like us cope with a world like the one we live in.

3.4 Summary

Premise semantics (3.1), conditional probability analyses (§3.2) and structural equation analyses (§3.3) all aim to analyze counterfactuals by focusing on certain relations between facts, rather than similarities between worlds. These accounts make clearer and more accurate predictions about particular counterfactuals in context than similarity analyses. But, ultimately, both premise semantics and conditional probability analyses had to incorporate causal dependence into their theories. Structural equation analyses do this from the start, and improve further on the predictions of premise semantics and conditional probability analyses. Another strength of this analysis is that it integrates elegantly into the broader applications of counterfactuals in theories of rationality, mental representation, causation and scientific explanation surveyed in §1.1. There is still rapid development of structural equation analyses, though, so it is too early to say where the analysis will stabilize, or how it will fair under thorough critical examination.

4 Conclusion

Philosophers, linguists and psychologists remain fiercely divided on how to best understand counterfactuals. Rightly so. They are at the center of questions of deep human interest (§1). The renaissance on this topic in the 70s and 80s focused on addressing certain semantic puzzles and capturing the logic of counterfactuals (§2). From this seminal literature, similarity analyses (Lewis 1973b; Stalnaker 1968) have enjoyed the most widespread popularity in philosophy (§2.3). But the logical debate between similarity and strict analyses is still raging, and strict analyses provide a viable logical alternative (§2.4). Criticisms of these logical analyses have focused recent debates on our intuitions about particular utterances of counterfactuals in particular contexts. Structural equation analyses (§3.3) have emerged as a particularly prominent alternative to similarity and strict analyses, which claims to improve on both in significant respects. These analyses are now being actively developed by philosophers, linguists, psychologists and computer scientists.

A Indicative and Subjunctive Conditionals

Historically, many philosophers have been tempted to assume that indicatives and subjunctives involve entirely different conditional connectives with related but substantially different meanings (Lewis 1973b; Gibbard 1981; Jackson 1987; Bennett 2003). This may be justifiable as an analytic convenience: one can use it to focus, as we are here, on two different complex constructions involving *if*, tense, aspect and modality, e.g. simple past conditionals vs. past perfect *would*-conditionals. But at a fully detailed level of analysis, there is no lexical ambiguity of *if*'s. There are just different effects produced by the elements with which *if* interacts. It might even turn out that there is no accurate binary semantic distinction between types of conditionals (Dudman 1988).[45] This is just to accept a minimal version of compositionality, the fact that different conditionals contain crucially different components and the fact that *if* is not

[45] A moment's thought reveals the staggering complexity here. For simple subject-predicate sentences there are at least 12 tense-aspect combinations. So even ignoring conditionals containing conditionals, there are at least 144 combinations to investigate in a conditional sentence that combines two sentences. The most comprehensive study for English is Declerck & Reed (2001: §5.7.5), which finds 9 importantly distinct tense combinations in counterfactuals — and this excludes many variants with modals or special verb forms in antecedent and consequent.

homonymous.[46]

Recent work in compositional semantics and pragmatics has offered sophisticated and connected explanations of the differences between various conditionals. Of most relevance here is the work on the difference between simple past and past perfect *would* conditionals. Even this subregion of the literature is too rich to fully detail here, but the main issue and positions will be summarized.

As in (5) and (6), simple past conditionals ('indicatives') do not admit of a counterfactual use, but past perfect *would* conditionals ('subjunctives') do. Stalnaker (1975) proposed that indicative antecedents evoke possibilities compatible with what's being assumed in the discourse, while subjunctives antecedents signal that no such assumption is being made. But this does not answer the key question: how can this difference be linked to the different morphology found in the two constructions? The two leading hypotheses begin with the observation that past tense marking in subjunctives does not receive its expected interpretation. The antecedent may be coherently supplemented with *tomorrow* to yield an antecedent that concerns a possibly counterfactual *future* event.[47]

(52) Bob died yesterday. If he had died **tomorrow** instead, he would have been 98 years old.

As evidenced by (54), this is not possible for a genuinely past tense reading of *Bob had died*, like that in (53).

(53) Yesterday I went to the Black Lodge. By the time I got there, Bob had died, but Cooper hadn't.

(54) I will go to the Black Lodge tomorrow. # By the time I get there, Bob had died, but Cooper hadn't.

This phenomenon does not occur with indicatives, and some have highlighted it as the dividing line between indicatives and subjunctives (Ippolito 2013: 2).[48]

(55) # If Bob danced tomorrow, Leland danced tomorrow.

This observation has generated two hypotheses. The first is:

Past as Remote Modality The past tense in subjunctives serves a modal function rather than a temporal one: it signals that the possibility described by the antecedent is not assumed to be among those regarded as actual in the discourse. (Isard 1974; Lyons 1977; Palmer 1986; Iatridou 2000; Schulz 2007, 2014; Starr 2014)

[46] If the *if*'s involved were mere homonyms then all conditionals should admit of both interpretations, which they do not. And, we should not expect to find the same particles used for the two constructions across unrelated languages, but we do.

[47] There are parallel examples for *were* and simple past tense subjunctives (Iatridou 2000).

[48] von Fintel (1999: §1) adds the requirement that the consequent have an overt modal like *might* or *would*. This definition of the indicative/subjunctive distinction is quite useful in English, but there is a question whether it is sufficiently cross-linguistically stable. For example, Bittner (2011) considers conditionals in the tenseless language Kalaallisut, where counterfactuals do not bear any morphological affinity with past tense. There, counterfactuals look just like indicatives except for the inclusion of a remote modality suffix *-galuar* in both antecedent and consequent.

There is actually some disagreement about the exact content of the Past as Remote Modality hypothesis, but the above is the weakest version.[49] This neatly explains why subjunctives can have a counterfactual use, but needn't be counterfactual. It also explains the intuition that subjunctives are logically stronger than indicatives: if it is appropriate to assert both, then the indicative follows from the subjunctive. It also generalizes more easily to languages which use a form for subjunctives that is unrelated to past tense. Schulz (2007, 2014) and Starr (2014) aim to address a gap in Past as Remote Modality analyses: they are not typically presented in a rigorous formal semantics. Past as Remote Modality approaches also inherit the responsibility of explaining why past tense would take on this modal meaning in these contexts. Iatridou (2000) proposes that remote modality and past tense have a common semantic feature: they express something about a topic that is remote from the perspective of the current discourse. This leaves room for a theory on which there is a single meaning that covers both uses, but that is refined to the modal use in conditional contexts. Schulz (2014) develops a polysemous account where the past tense in subjunctives contributes a distinct modal operator arises from the diachronic process of re-categorization.[50]

The alternative hypothesis about past tense in subjunctives is this:

Past Modality The past tense in subjunctives modifies the modality rather than the time at which antecedent and consequent hold: it conveys what was possible/necessary in the past (Adams 1976; Skyrms 1981; Tadeschi 1981; Dudman 1984a,b; Arregui 2007, 2009; Ippolito 2006, 2008, 2013; Khoo 2015)

Past Modality approaches inherit no such responsibility because they find room for a normal past tense meaning by allowing for it to operate on the modality expressed, rather than the events described by the conditional. This approach begins with the idea that what's possible is a function of time: what's possible now is very different than what was possible at a prior point in history. For example, Tadeschi (1981) thinks of each event as a branch point that distinguishes one world from another. At the present time in the actual world, certain branches are open, but certain branches had to have been taken to have ended up here, now. The past tense in a subjunctive can then move one back to any salient earlier branch point and express generalizations about the branches open from there. However, from that past time, there will still be facts about what happens tomorrow. So, while the past tense changes the sense of possibility at

[49] This is the version preferred by Schulz (2007, 2014) and Starr (2014), while Iatridou (2000) prefers a stronger position: subjunctive antecedents evoke a scenario which is assumed to be partly incompatible with what is being assumed about the actual world in the discourse. The most comprehensive study of the possibilities here is von Fintel (1999), although that study is more broadly construed so as to be applicable to Past Modality approaches: what do indicatives presuppose about the possibilities they describe?

[50] Roughly, re-categorization is the adaptation of an old form to a new use which exploits some similarity between the uses. Since the uses are not identical, a new convention evolves and hence a new morpheme is born.

play, adverbs like *tomorrow* can still refer to our current future, they just do so from a shifted modal perspective.

On the Past Modality approach, counterfactual uses are made possible by the assumption that something which is now impossible may have once been possible. Of course, non-counterfactual uses are also accommodated: one may want to describe what was possible at a past point in our timeline. Another crucial detail for Past Modality approaches is the treatment of simple past indicative conditionals. They are analyzed as a species of epistemic modality about a past event rather than historical modality (Tadeschi 1981; Arregui 2007; Ippolito 2013). This is needed to explain the truth-conditional differences between subjunctives and indicatives, e.g. (3)/(4).

Past Modality theories require the logical form of subjunctives to depart from its surface form: tense must take scope over the conditional. Various plausible ways of working this out have been explored (Tadeschi 1981; Arregui 2007; Ippolito 2013). A more significant concern arises from hindsight counterfactuals such as (56).

(56) *You have been offered a $1 bet on the outcome of a fair coin toss. You decline. The coin is flipped and comes up heads.* If you had bet heads, you would have won $1.

Intuitively, the counterfactual is true. But, it is hard to see how this can be true on a past modality approach (Barker 1998; Edgington 2004; Schulz 2007). If one goes back to a time before the coin flip, it seems unlikely that on even most of the branches where you bet, the coin toss comes out the way it actually does. But then you wouldn't have won $1, and the counterfactual would be false. Ippolito (2013: §3.6) offers a Past Modality approach designed to address this issue. Another potential issue arises with epistemic uses of counterfactuals like the following one from Hansson (1989).[51]

> *Suppose that one night you approach a small town of which you know that it has exactly two snack bars. Just before entering the town you meet a man eating a hamburger. You have good reason to accept the following conditional*:
>
> (57) If snack bar A is closed, snack bar B is open
>
> Suppose now that after entering the town and seeing the man with the burger, you see that A is in fact open. I comment on how lucky we are that this one is open. You temper my enthusiasm by reminding me:
>
> (58) If snack bar A had been closed, snack bar B would have been open

[51] See also the King Ludwig example in Kratzer (1989: 640). For a detailed discussion of these examples see Schulz (2007: §5.6).

(58) appears to be a counterfactual conditional about the past, but which expresses something about our present epistemic state: there's no temptation to read it as claiming that the closure of snack bar B caused A to be open. This challenges the assumption Past Modality approaches rely on to distinguish subjunctives from simple past indicatives: only simple past indicatives express something about our present information about past events. Incidentally, these epistemic counterfactuals are also a difficult case for many philosophical theories which aim to identify the indicative/subjunctive distinction with the epistemic/metaphysical modality distinction, e.g., Weatherson (2001). For the distinction between epistemic and metaphysical modality, see the entry Varieties of Modality.

B Formal Constraints on Similarity

In addition to the limit assumption, here is a list of the formal constraints on similarity that have been proposed, where $p, q \subseteq W$ and $w \in W$:

(a)	$f(w,p) \subseteq p$	**success**
(b)	$f(w,p) = \{w\}$, if $w \in p$	**strong centering**
(c)	$f(w,p) \subseteq q$ & $f(w,q) \subseteq p \implies f(w,p) = f(w,q)$	**uniformity**
(d)	$f(w,p)$ contains *at most* one world	**uniqueness**

Table 5: Candidate Constraints on Selection Functions

While success is discussed in the main entry, (b)–(d) are discussed below, followed by the limit assumption.

B.1 Strong Centering

Strong centering is motivated, in part, by the intuitive concept of similarity: if w is already a p-world, then the p-world most similar to w is w itself. But logical concerns also motivate its inclusion. Something in the vicinity of strong centering is needed to validate modus ponens ($\phi > \psi, \phi \vDash \psi$), which underwrites reasoning like (59).

(59) a. If George were caught, he would face years of prison.
b. Actually, George did get caught.
c. In that case, he must be facing years of prison.

Similarly, strong centering validates the principle that a subjunctive conditional is false if its antecedent is true and consequent false: $\phi \wedge \neg\psi \vDash \neg(\phi > \psi)$. However, **weak centering** would suffice for both: $w \in f(w,p)$ if $w \in p$. The difference is that strong centering, but not weak centering, validates:

Conjunction Conditionalization $\phi \wedge \psi \vDash \phi > \psi$

This principle was only meekly promoted by Lewis (1973b: §1.7), and has attracted objections leading some to adopt a similarity analysis with only weak centering (e.g. Bennett 1974: 387-8). However, Walters & Williams (2013) provide a very thorough defense of Conjunction Conditionalization and offer a helpful survey of the objections to it.

B.2 Uniformity

The uniformity constraint is somewhat more difficult to state intuitively.[52] But its primary purpose is to validate SSE (Stalnaker 1984: 130), which allows one to substitute subjunctive equivalents in the antecedent position. It also bears noting that *limited* forms of Transitivity and Antecedent Strengthening follow directly from SSE:[53]

Substitution of Subjunctive Equivalents (SSE)

$$\phi_1 > \phi_2, \phi_2 > \phi_1, \phi_1 > \psi \vDash \phi_2 > \psi$$

Limited Transitivity (LT)

$$\phi_1 > \phi_2, (\phi_1 \wedge \phi_2) > \psi \vDash \phi_1 > \psi$$

Limited Antecedent Strengthening (LAS)

$$\phi_1 > \phi_2, \neg(\phi_1 > \neg\psi) \vDash (\phi_1 \wedge \phi_2) > \psi$$

The intuitive appeal of SSE is fairly clear:

(60) a. If Simone had drummed, Jean-Paul would have danced.
b. If Jean-Paul had danced, Simone would have drummed.
c. If Simone had drummed, Claude would have stayed for another drink.
d. So, if Jean-Paul had danced, Claude would have stayed for another drink.

LT and LAS are also important since they provide one way the similarity analysis could explain the fact that some instances of Transitivity and Antecedent Strengthening sound compelling. If the similarity theorist can motivate the claim that the premises of these instances are being interpreted in the manner of LAS rather than Antecedent Strengthening, then these cases can be explained.

Pollock (1976, 1981: 254) rejects uniformity and endorses an even weaker logic. Pollock (1981: 254) does so on the basis of a counterexample to LAS. In the example, there is a circuit where three switches (S_1, S_2, S_3) control two lights (L_1, L_2). L_1 comes on either when S_1 is up or when S_2 and S_3 are both up. L_2 comes on when S_1 is up or when S_2 is up. (S_3 is not connected to L_2

[52] Stalnaker (1984: 129) puts it this way: “This is the requirement that if one possible world is selected over another relative to one antecedent, then it must be favored relative to any antecedent for which both are eligible.” But it takes some work to see how this is a paraphrase of uniformity.

[53] For LT, one can get from the second premise to the conclusion by SSE if $(\phi_1 \wedge \phi_2) > \phi_1$ and $\phi_1 > (\phi_1 \wedge \phi_2)$ can be established. The former is a logical truth, and the latter follows from the first premise. For LAS, one can get from the second premise to the conclusion by SSE, and as before $(\phi_1 \wedge \phi_2) > \phi_1$ and $\phi_1 > (\phi_1 \wedge \phi_2)$.

in any way.) Now suppose that all three switches are down and both lights are off. Both (61a) and (61b) are intuitively true. What about (61c)?

(61) a. S_3 would (still) be down if L_2 were on. $\mathsf{L}_2 > \neg \mathsf{S}_3$
b. It's not true that if L_2 were on, L_1 would be off. (L_2 might be on because S_1 is up.) $\neg(\mathsf{L}_2 > \neg \mathsf{L}_1)$
c. L_1 would be on if L_2 were on and S_3 were down. $(\mathsf{L}_2 \wedge \neg \mathsf{S}_3) > \mathsf{L}_1$

S_3 would (still) be down if L_2 and L_1 were on. Pollock (1981) contends that (61c) makes the wrong prediction in a scenario where S_1 and S_3 stay down but S_2 is flipped up. Then L_2 will be on and S_3 will be down, but L_1 will be off. Pollock (1981: §2) maintains that this failure of uniformity motivates an account in terms of minimal change instead of one in terms of maximal similarity. This counterexample has received little attention, perhaps because it is sufficiently complex to make intuitions less clear. It is also worth noting that rejecting uniformity makes it harder to explain why some instances of antecedent monotonic patterns sound compelling.

B.3 Uniqueness

According to uniqueness, there is always a unique most similar ϕ-world when evaluating $\phi > \psi$. For such a simple principle, uniqueness has stimulated a surprisingly complex debate, beginning with Stalnaker's (1968: 46) endorsement and Lewis' (1973b: §3.4) rejection of it. This debate centers on two issues: the fact that Uniqueness (together with the limit assumption discussed below) entails Conditional Excluded Middle and the fact that Uniqueness impacts one's analysis of subjunctive conditionals containing *might* in the consequent. Consider first Conditional Excluded Middle and two consequences, CN and CD:[54]

Conditional Excluded Middle (CEM)

$$\vDash (\phi > \psi) \vee (\phi > \neg\psi)$$

Conditional Negation (CN)

$$\Diamond\phi \wedge \neg(\phi > \psi) \dashv\vDash \Diamond\phi \wedge \phi > \neg\psi$$

Consequent Distribution (CD)

$$\phi > (\psi_1 \vee \psi_2) \vDash (\phi > \psi_1) \vee (\phi > \psi_2)$$

Uniqueness leads to CEM, since if there is a unique most similar ϕ-world it is either a ψ-world — in which case the left disjunct is true — or it is a $\neg\psi$-world — in which case the right disjunct is true. A natural concern, voiced by Lewis (1973b: §3.4), is that there might be a tie among a ψ-world and a $\neg\psi$-world for being the most similar ϕ-world.

[54] CN requires adding $\Diamond\phi$ to cover the case where ϕ is contradictory and thus both $\phi > \psi$ and $\phi > \neg\psi$ are vacuously true and $\neg(\phi > \psi)$ false. $\Diamond\phi$ is given a similarity analysis as well, where it's true just in case $f(w, \llbracket\phi\rrbracket_v^f) \neq \varnothing$.

Consider the antecedent *If I were older…* What is my age in the unique most similar world where I am older? One might be tempted to deny both (62a) and (62b), instead affirming (62c).

(62) a. If I were older, I would be 35.
b. If I were older, I wouldn't be 35.
c. If I were older, I might be 35. And, if I were older, I might not be 35.

It seems appealing that all worlds where I am older than my present age, say 32, are tied. The advocate of uniqueness has two problems here. First, they cannot deny both (62a) and (62b), since by CEM at least one of them is true. Second, it is unclear what they can say about the meaning of (62c). It is tempting to say, as Lewis (1973b: §1.5) does, that a *might* subjunctive is true when *some* of the most similar antecedent worlds are consequent worlds. But if there is a unique most similar antecedent world, the difference between requiring *all* for *would* subjunctives and *some* for *might* subjunctives collapses.

Stalnaker (1981, 1984: Ch.7) responds to both challenges. He begins by highlighting favorable data for CEM, contending, as Lewis (1973b: §3.4) grants, that (63c) sounds inconsistent with (63a) and (63b).

(63) a. It's not true that if I were older, I would be 35.
$\neg(\mathsf{O} > \mathsf{T})$
b. And, it's not true that if I were older, I wouldn't be 35.
$\neg(\mathsf{O} > \neg\mathsf{T})$
c. But, if I were older, I either would or would not be 35.
$\neg(\mathsf{O} > (\mathsf{T} \vee \neg\mathsf{T}))$

But (63c) is only inconsistent with (63a) and (63b) when CEM is assumed. This is easy to see from the fact that CEM entails CD, and by CD (63c) entails $(\mathsf{O} > \mathsf{T}) \vee (\mathsf{O} > \neg\mathsf{T})$, which is clearly inconsistent with (63a) and (63b). And, yet, CEM also conflicts with our intuition that (63a) and (63b) are consistent with each other. Stalnaker (1981, 1984: Ch.7) proposes to treat such cases as vagueness in the similarity of worlds, on par with the vagueness of color gradients where one might be tempted to deny both that a certain patch is red and that it is not red. The basic idea is that there is indeterminacy in the selection function being used and under one resolution (63a) is true while under another (63b) is true. Stalnaker (1981, 1984: Ch.7) sketches a view along these lines using the supervaluational (Van Fraassen 1966) approach to vagueness. The remaining challenge, then, is to formulate an approach to *might* subjunctives like (62c) which is compatible with uniqueness. Stalnaker (1981, 1984: Ch.7) develops just such an account.

For the advocate of uniqueness, it is difficult to treat *might* subjunctives as involving a subjunctive conditional whose consequent is an existentially quantified modal. Since the antecedent delivers a single most similar world, it is not possible to truth-conditionally distinguish a universal from an existential quantifier over the antecedent worlds. Thus, a *might* or *could* and a *would* will come to the same thing.[55] As a result, Stalnaker (1981, 1984: Ch.7) pursues an analysis where *might* takes scope over the entire subjunctive conditional and expresses a kind of epistemic possibility. Stalnaker (1984: 143-4) motivates this by observing that (64b) seems equivalent to (64a).

(64) a. If John had been invited, he might have come to the party.
b. It might be that if John had been invited, he would have come to the party.

Stalnaker (1984: 144-6) goes on to present examples to support this analysis and compare it favorably with Lewis' (1973b: §1.5).

The debate regarding CEM has continued to attract arguments on each side. Bennett (2003: §§72,73,76) presents a sustained critique of CEM, while Williams (2010) responds on its behalf. von Fintel & Iatridou (2002) examine quantified conditionals and suggest that a semantics which validates CEM is needed. Klinedinst (2011) develops this argument further against intervening disputes. Swanson (2012) shows that the supervaluationist approach can also be applied to handle failures of the limit assumption, thereby allowing for a theory that validates CEM despite having mechanisms for treating failures of both uniqueness and the limit assumption.

There is another issue raised by the discussion of *might* conditionals which has attracted much attention. There is an imperfect match between the representation language being used and English. The connective '>' builds in *would*, and *might* subjunctives contain no *would*. Rather than being analyzed as a *might* scoping over a *would* conditional (Stalnaker 1981, 1984: Ch.7) or a distinct conditional connective (Lewis 1973b: §1.5), it would be more accurate to have an analysis that separates the contribution of conditionals in general from the modals that interact with them. This is not an idle linguistic issue. Neither (Stalnaker 1984: Ch.7) nor (Lewis 1973b: §1.5) can capture examples like (65).

(65) If John had come to the party, he would have had a drink and he might have liked it.

Scoping the *might* over the conditional would erase the crucial distinction made in the consequent: John definitely would have had a drink, but there's only a chance that he would have liked it.[56] These issues can only be addressed by a more serious engagement with the research discussed in the supplement Indicative and Subjunctive Conditionals, especially Kratzer's (1981a; 1991) general approach to modality and conditionals.

[55] Evaluated in w_0, $\mathsf{O} > \mathsf{Could}(\mathsf{T})$ says that $\mathsf{Could}(\mathsf{T})$ is true in the O-world most similar to w_0, call it w_1. Either w_1 is a T-world or a $\neg\mathsf{T}$-world. If w_1 is a T-world, then $\mathsf{O} > \mathsf{T}$ is true. If w_1 is a $\neg\mathsf{T}$-world, then $\mathsf{O} > \neg\mathsf{T}$ is true.

[56] Though it is syntactically suspect, one might try to analyze it as a conjunction of conditionals. But this distorts the anaphoric relations: *if John had come to the party, he would have had a drink and it might be that if he had come to the party, he would have liked it.* On this re-analysis, *it*'s dominant construal is *the party*, unlike (65).

B.4 The Limit Assumption

The basic idea of the limit assumption is that there is are most similar antecedent worlds:

Limit Assumption

> As one proceeds to ϕ-worlds more and more similar to w, one hits a limit and cannot get to a ϕ-world any more similar to w.

While Stalnaker (1968, 1981) accepts this assumption, Lewis (1973b: 20) rejects it. Lewis (1973b: 20) provides the following rationale. Consider this 1 inch line:

Suppose, counterfactually, that this line were more than an inch long. How long is it in the world most similar to our own? For any length of line and corresponding world $1 + x$, there is a real number $y < x$ such that the line is $1 + y$ in some other world. This other world is yet more similar to our own, since the line is closer to its actual length. Lewis (1973b: 20) grants that this is not a decisive objection, but suggests that it is better to formulate a theory of counterfactuals which does not essentially depend on the limit assumption to operate. Lewis (1973c: 423-4) prefers an analysis according to which $\phi > \psi$ is true in w just in case some $\phi \wedge \psi$-world is more similar to w than any $\phi \wedge \neg\psi$-world, if there are any ϕ-worlds. This analysis can be formalized in terms of a three-place comparative similarity relation over worlds that is subject to constraints analogous to those devised for selection functions (Lewis 1973b: §2.3).

Stalnaker (1981, 1984: 141-2) replies that in most contexts not every minute difference will count towards the similarity of antecedent worlds. Stalnaker (1968) and Lewis (1973b) share the commitment that the respects which matter for similarity are radically context-sensitive. Thus, in cases like the line, they both seem to agree that there will be a threshold below which differences don't matter. Stalnaker (1981, 1984: 141-2) contends that this threshold provides the limit needed to safely make the limit assumption. In the context of the line, the threshold is likely the smallest unit of measure which people can be assumed to care about, e.g. a millimeter. As for contexts where *every* difference matters, Stalnaker (1984: 141-2) says that it is inappropriate to use imprecise antecedents like *if this line were more than an inch long*, much as it is inappropriate to use a definite description like *the shortest lines longer than an inch*. If correct, this response eliminates the need to countenance failures of the limit assumption. However, Swanson (2012) shows that even if one must countenance such failures, there is a supervaluationist method — extending considerably Stalnaker's (1981) — for treating them within a theory that makes the limit assumption.

Pollock (1976: 20) and Herzberger (1979) also highlight one advantage of making the limit assumption. All theories on offer agree that if $\phi_1 > \phi_2$ is true and $\phi_2 \vDash \psi$ then $\phi_1 > \psi$ is true. And, they agree that if $\phi_1 > \phi_2, \ldots, \phi_1 > \phi_n$ are true and $\phi_2, \ldots, \phi_n \vDash \psi$ then $\phi_1 > \psi$ is true. But what about the more general version: where $\Gamma = \{\phi_2, \phi_3, \ldots\}$, if $\phi_1 > \phi_2, \phi_1 > \phi_3, \ldots$ are true and $\Gamma \vDash \psi$ then $\phi_1 > \psi$ is true? The more general version only holds when the limit assumption is made. The appeal of this general consequence principle, plus the above resources, mean that the limit assumption is often safely made in discussion of counterfactuals. If for nothing else, this assumption often simplifies already complex formal definitions. However, Kaufmann (2017) demonstrated a rather important caveat: formalizing the Limit Assumption reveals it to be not one assumption, but a family of assumptions which must be stated differently depending on other features of the formalization. While the formulations of it given in Lewis (1973b) are accurate given certain assumptions about the underlying ordering relations, different formulations are required when those assumptions are changed.

References

Adams, E (1965). 'The Logic of Conditionals.' *Inquiry*, **8**: 166–197.

Adams, EW (1970). 'Subjunctive and Indicative Conditionals.' *Foundations of Language*, **6(1)**: pp. 89–94. URL http://www.jstor.org/stable/25000429.

Adams, EW (1975). *The Logic of Conditionals*. Dordrecht: D. Reidel.

Adams, EW (1976). 'Prior Probabilities and Counterfactual Conditionals.' In W Harper & C Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, vol. 6a of *The University of Western Ontario Series in Philosophy of Science*, 1–21. Springer Netherlands. URL http://dx.doi.org/10.1007/978-94-010-1853-1_1.

Alastair, W (2017). 'Metaphysical Causation.' *Noûs*. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/nous.12190.

Alonso-Ovalle, L (2009). 'Counterfactuals, Correlatives, and Disjunction.' *Linguistics and Philosophy*, **32(2)**: 207–244. URL http://dx.doi.org/10.1007/s10988-009-9059-0.

Alquist, JL, Ainsworth, SE, Baumeister, RF, Daly, M & Stillman, TF (2015). 'The Making of Might-Have-Beens: Effects of Free Will Belief on Counterfactual Thinking.' *Personality and Social Psychology Bulletin*, **41(2)**: 268–283. URL https://doi.org/10.1177/0146167214563673.

Anderson, AR (1951). 'A Note on Subjunctive and Counterfactual Conditionals.' *Analysis*, **12(2)**: pp. 35–38. URL http://www.jstor.org/stable/3327037.

Arregui, A (2007). 'When aspect matters: the case of would-conditionals.' *Natural Language Semantics*, **15**: 221–264. URL http://dx.doi.org/10.1007/s11050-007-9019-6.

Arregui, A (2009). 'On similarity in counterfactuals.' *Linguistics and Philosophy*, **32**: 245–278. URL http://dx.doi.org/10.1007/s10988-009-9060-7.

Barker, SJ (1998). 'Predetermination and tense probabilism.' *Analysis*, **58(4)**: 290–296. URL http://analysis.oxfordjournals.org/content/58/4/290.

Bennett, J (1974). 'Counterfactuals and Possible Worlds.' *Canadian Journal of Philosophy*, **4(2)**: 381–402. URL http://www.jstor.org/stable/40230514.

Bennett, J (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bennett, K (2017). *Making Things Up*. New York: Oxford University Press.

Bittner, M (2011). 'Time and Modality without Tenses or Modals.' In R Musan & M Rathers (eds.), *Tense Across Languages*, 147–188. Tübingen: Niemeyer. URL http://semanticsarchive.net/Archive/zliYmQxY/bittner11_tam.pdf.

Bobzien, S (2011). 'Dialectical School.' In EN Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, fall 2011 edn. URL http://plato.stanford.edu/archives/fall2011/entries/dialectical-school/.

Bowie, GL (1979). 'The Similarity Approach to Counterfactuals: Some Problems.' *Noûs*, **13(4)**: pp. 477–498. URL http://www.jstor.org/stable/2215340.

Brée, D (1982). 'Counterfactuals and Causality.' *Journal of Semantics*, **1(2)**: 147–185. URL http://dx.doi.org/10.1093/jos/1.2.147.

Briggs, RA (2012). 'Interventionist counterfactuals.' *Philosophical Studies*, **160(1)**: 139–166. URL http://dx.doi.org/10.1007/s11098-012-9908-5.

Byrne, RM (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

Byrne, RM (2016). 'Counterfactual Thought.' *Annual Review of Psychology*, **67(1)**: 135–157. URL https://doi.org/10.1146/annurev-psych-122414-033249.

Carnap, R (1948). *Introduction to Semantics*. Cambridge, MA: Harvard University Press.

Carnap, R (1956). *Meaning and Necessity*. 2 edn. Chicago: Chicago University Press. (First edition published in 1947.).

Champollion, L, Ciardelli, I & Zhang, L (2016). 'Breaking de Morgan's Law in Counterfactual Antecedents.' In M Moroney, CR Little, J Collard & D Burgdorf (eds.), *Proceedings from Semantics and Linguistic Theory (SALT) 26*, 304–324. Ithaca, NY: CLC Publications.

Chater, N, Oaksford, M, Hahn, U & Heit, E (2010). 'Bayesian models of cognition.' *Wiley Interdisciplinary Reviews: Cognitive Science*, **1(6)**: 811–823. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.79.

Chisholm, RM (1955). 'Law Statements and Counterfactual Inference.' *Analysis*, **15(5)**: 97–105. URL http://www.jstor.org/stable/3326359.

Ciardelli, I, Zhang, L & Champollion, L (2018). 'Two switches in the theory of counterfactuals.' *Linguistics and Philosophy*. URL https://doi.org/10.1007/s10988-018-9232-4.

Cohen, J & Meskin, A (2006). 'An objective counterfactual theory of information.' *Australasian Journal of Philosophy*, **84(3)**: 333–352. URL https://doi.org/10.1080/00048400600895821.

Copeland, BJ (2002). 'The Genesis of Possible Worlds Semantics.' *Journal of Philosophical Logic*, **31(2)**: 99–137. URL http://dx.doi.org/10.1023/A%3A1015273407895.

Costello, T & McCarthy, J (1999). 'Useful Counterfactuals.' *Linköping Electronic Articles in Computer and Information Science*, **4(12)**: 1–24. URL http://www.ep.liu.se/ea/cis/1999/012/.

Cresswell, MJ & Hughes, G (1996). *A New Introduction to Modal Logic*. London: Routledge.

Daniels, CB & Freeman, JB (1980). 'An analysis of the subjunctive conditional.' *Notre Dame J. Formal Logic*, **21(4)**: 639–655. URL https://doi.org/10.1305/ndjfl/1093883247.

Declerck, R & Reed, S (2001). *Conditionals: A Comprehensive Emprical Analysis*, vol. 37 of *Topics in English Linguistics*. New York: De Gruyter Mouton.

Dretske, F (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Dretske, F (2002). 'A Recipe for Thought.' In DJ Chalmers (ed.), *Philosophy of Mind: Contemporary and Classical Readings*, chap. 46, 491–9. New York: Oxford University Press.

Dretske, F (2011). 'Information-Theoretic Semantics.' In B McLaughlin, A Beckermann & S Walter (eds.), *The Oxford Handbook of Philosophy of Mind*, 381–393. New York: Oxford University Press.

Dretske, FI (1981). *Knowledge and the Flow of Information*. Cambridge, Massachusetts: The MIT Press.

Dudman, VC (1984a). 'Conditional Interpretations of 'If' Sentences.' *Australian Journal of Linguistics*, **4(2)**: 143–204.

Dudman, VH (1984b). 'Parsing 'If'-Sentences.' *Analysis*, **44(4)**: 145–153. URL http://analysis.oxfordjournals.org/content/44/4/145.

Dudman, VH (1988). 'Indicative and Subjunctive.' *Analysis*, **48(3)**: 113–122. URL http://analysis.oxfordjournals.org/content/48/3/113.2.

Edgington, D (2003). 'What If? Questions About Conditionals.' *Mind & Language*, **18(4)**: 380–401.

Edgington, D (2004). 'Counterfactuals and the benefit of hindsight.' In PDP Noordhof (ed.), *Cause and Chance: Causation in an Indeterministic World*, 12–27. New York: Routledge.

Fine, K (1975). 'Review of Lewis' *Counterfactuals*.' *Mind*, **84**: 451–8.

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

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

von Fintel, K (1999). 'The Presupposition of Subjunctive Conditionals.' In U Sauerland & O Percus (eds.), *The Interpretive Tract*, vol. MIT Working Papers in Linguistics 25, 29–44. Cambridge, MA: MITWPL. URL http://mit.edu/fintel/www/subjunctive.pdf.

von Fintel, K (2001). 'Counterfactuals in a Dynamic Context.' In M Kenstowicz (ed.), *Ken Hale: a Life in Language*, 123–152. Cambridge, Massachusetts: The MIT Press. URL http://mit.edu/fintel/www/conditional.pdf.

von Fintel, K (2012). 'Subjunctive Conditionals.' In G Russell & DG Fara (eds.), *The Routledge Companion to Philosophy of Language*, 466–477. New York: Routledge. URL http://mit.edu/fintel/fintel-2012-subjunctives.pdf.

von Fintel, K & Iatridou, S (2002). 'If and When *If*-Clauses Can Restrict Quantifiers.' Ms. MIT, Department of Linguistics and Philosophy, URL http://web.mit.edu/fintel/www/lpw.mich.pdf.

Fisher, T (2017a). 'Causal counterfactuals are not interventionist counterfactuals.' *Synthese*, **194(12)**: 4935–4957. URL https://doi.org/10.1007/s11229-016-1183-0.

Fisher, T (2017b). 'Counterlegal dependence and causation's arrows: causal models for backtrackers and counterlegals.' *Synthese*, **194(12)**: 4983–5003. URL https://doi.org/10.1007/s11229-016-1189-7.

Fodor, JA (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, Massachusetts: The MIT Press.

Fodor, JA (ed.) (1990). *A Theory of Content and Other Essays*. Cambrudge, Massachusetts: The MIT Press.

Frege, G (1893). *Grundgesetze der Arithmetik, begriffsschriftlich abgeleitet, Vol. 1*. 1st edn. Jena: H. Pohle.

Galinsky, AD, Liljenquist, KA, Kray, LJ & Roese, NJ (2005). 'Finding meaning from mutability: Making sense and deriving significance through counterfactual thinking.' In *The Psychology of Counterfactual Thinking*, 110–127. New York: Routledge.

Gamut, LTF (1991). *Logic, Language and Meaning: Intensional Logic and Logical Grammar*, vol. 2. The University of Chicago Press.

Gärdenfors, P (1978). 'Conditionals and Changes of Belief.' In I Niiniluoto & R Tuomela (eds.), *The Logic and Epistemology of Scientific Belief*. Amsterdam: North-Holland.

Gärdenfors, P (1982). 'Imaging and Conditionalization.' *Journal of Philosophy*, **79(12)**: 747–760.

Gibbard, AF (1981). 'Two Recent Theories of Conditionals.' In WL Harper, RC Stalnaker & G Pearce (eds.), *Ifs: Conditionals, Beliefs, Decision, Chance, Time*, 211–247. Dordrecht: D. Reidel.

Gibbard, AF & Harper, WL (1978). 'Counterfactuals and Two Kinds of Expected Utility.' In C Hooker, JJ Leach & E McClennen (eds.), *Foundations and Applications of Decision Theory*, 125–162. Dordrecht: D. Reidel.

Gillies, A (2007). 'Counterfactual Scorekeeping.' *Linguistics & Philosophy*, **30(3)**: 329–360. URL http://rci.rutgers.edu/~thony/counterfactualscorekeeping_landp.pdf.

Gillies, A (2012). 'Indicative Conditionals.' In G Russell & DG Fara (eds.), *The Routledge Companion to Philosophy of Language*, 449–465. New York: Routledge.

Ginsburg, ML (1985). 'Counterfactuals.' In A Joshi (ed.), *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 80–86. Los Altos, California: Morgan Kaufmann.

Glymour, C (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge: Cambridge University.

Goodman, N (1947). 'The Problem of Counterfactual Conditionals.' *The Journal of Philosophy*, **44**: 113–118.

Goodman, N (1954). *Fact, Fiction and Forecast*. Cambridge, MA: Harvard University Press.

Gopnik, A, Glymour, C, Sobel, DM, Schulz, LE, Kushnir, T & Danks, D (2004). 'A Theory of Causal Learning in Children: Causal Maps and Bayes Nets.' *Psychological Review*, **111(1)**: 3–32.

Gopnik, A & Tenenbaum, JB (2007). 'Bayesian networks, Bayesian learning and cognitive development.' *Developmental Science*, **10(3)**: 281–287. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7687.2007.00584.x.

Greenberg, G (2013). 'An Argument for the Strict Analysis of Subjunctive Conditionals.' Ms. UCLA, URL http://gjgreenberg.bol.ucla.edu/greenberg_strict.pdf.

Hájek, A (2014). 'Probabilities of counterfactuals and counterfactual probabilities.' *Journal of Applied Logic*, **12(3)**: 235 – 251. Special Issue on Combining Probability and Logic to Solve Philosophical Problems, URL http://www.sciencedirect.com/science/article/pii/S1570868314000160.

Halpern, J & Pearl, J (2005a). 'Causes and Explanations: A Structural-Model Approach. Part I: Causes.' *British Journal for Philosophy of Science*, **56**.

Halpern, J & Pearl, J (2005b). 'Causes and Explanations: A Structural-Model Approach. Part II: Explanations.' *British Journal for Philosophy of Science*, **56**: 889–911.

Hansson, SO (1989). 'New Operators for Theory Change.' *Theoria*, **55**: 114–132.

Harper, WL (1975). 'Rational Belief Change, Popper Functions and the Counterfactuals.' *Synthése*, **30**: 221–262.

Hawthorne, J (2005). 'Chance and Counterfactuals.' *Philosophy and Phenomenological Research*, **70(2)**: 396–8211.

Heintzelman, SJ, Christopher, J, Trent, J & King, LA (2013). 'Counterfactual thinking about one's birth enhances well-being judgments.' *The Journal of Positive Psychology*, **8(1)**: 44–49. URL https://doi.org/10.1080/17439760.2012.754925.

Herzberger, H (1979). 'Counterfactuals and Consistency.' *Journal of Philosophy*, **76(2)**: 83–88.

Hiddleston, E (2005). 'A Causal Theory of Counterfactuals.' *Noûs*, **39(4)**: 632–657. URL http://dx.doi.org/10.1111/j.0029-4624.2005.00542.x.

Hitchcock, C (2001). 'The Intransitivity of Causation Revealed by Equations and Graphs.' *The Journal of Philosophy*, **98(6)**: 273–299. URL http://www.jstor.org/stable/2678432.

Hitchcock, C (2007). 'Prevention, Preemption, and the Principle of Sufficient Reason.' *Philosophical Review*, **116(4)**: 495–532.

Horwich, P (1987). *Asymmetries in Time*. Cambridge, MA: MIT Press.

Iatridou, S (2000). 'The Grammatical Ingredients of Counterfactuality.' *Linguistic Inquiry*, **31(2)**: 231–270.

Ichikawa, J (2011). 'Quantifiers, Knowledge, and Counterfactuals.' *Philosophy and Phenomenological Research*, **82(2)**: 287–313. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1933-1592.2010.00427.x.

Ippolito, M (2006). 'Semantic Composition and Presupposition Projection in Subjunctive Conditionals.' *Linguistics and Philosophy*, **29(6)**: 631–672. URL http://dx.doi.org/10.1007/s10988-006-9006-2.

Ippolito, M (2008). 'Subjunctive Conditionals.' In A Grønn (ed.), *Proceedings of Sinn und Bedeutung 12*, 256–270. Oslo: Department of Literature, Area Studies and European Languages, University of Oslo. URL http://www.hf.uio.no/ilos/forskning/aktuelt/arrangementer/konferanser/2007/SuB12/proceedings/.

Ippolito, M (2013). *Subjunctive Conditionals: a linguistic analysis*. No. 65 in Linguistic Inquiry Monograph Series. Cambridge, MA: MIT Press.

Ippolito, M (2016). 'How similar is similar enough?' *Semantics and Pragmatics*, **9(6)**: 1–60.

Isard, S (1974). 'What Would You Have Done If...' *Theoretical Linguistics*, **1**: 233–55.

Jackson, F (1987). *Conditionals*. Oxford: Basil Blackwell.

Kahneman, D, Slovic, P & Tversky, A (eds.) (1982). *Judgement under Uncertainty: heuristics and biases*. Cambridge: Cambridge University Press.

Kant, I (1781/1787/1987). *Critique of Pure Reason*. Cambridge: Cambridge University Press.

Kaufmann, S (2005). 'Conditional Predictions.' *Linguistics and Philosophy*, **28(2)**: 181–231. URL http://dx.doi.org/10.1007/s10988-005-3731-9.

Kaufmann, S (2008). 'Conditionals Right and Left: Probabilities for the Whole Family.' *Journal of Philosophical Logic*, **38(1)**: 1–53. URL http://dx.doi.org/10.1007/s10992-008-9088-0.

Kaufmann, S (2013). 'Causal Premise Semantics.' *Cognitive Science*, **37(6)**: 1136–1170. URL http://dx.doi.org/10.1111/cogs.12063.

Kaufmann, S (2017). 'The Limit Assumption.' *Semantics and Pragmatics*, **10(18)**.

Khoo, J (2015). 'On Indicative and Subjunctive Conditionals.' *Philosophers' Imprint*, **15(32)**: 1–40. URL http://hdl.handle.net/2027/spo.3521354.0015.032.

Klinedinst, N (2011). 'Quantified Conditionals and Conditional Excluded Middle.' *Journal of Semantics*, **28(1)**: 149–170. URL http://jos.oxfordjournals.org/content/28/1/149.abstract.

Kment, B (2006). 'Counterfactuals and Explanation.' *Mind*, **115(458)**: 261–310. URL http://mind.oxfordjournals.org/content/115/458/261.

Kment, B (2014). *Modality and Explanatory Reasoning*. New York: Oxford University Press.

Koslicki, K (2016). 'Where grounding and causation part ways: comments on Schaffer.' *Philosophical Studies*, **173(1)**: 101–112. URL https://doi.org/10.1007/s11098-014-0436-3.

Kratzer, A (1981a). 'The Notional Category of Modality.' In HJ Eikmeyer & H Rieser (eds.), *Words, Worlds and Contexts*, 38–74. Berlin: Walter de Gruyter.

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

Kratzer, A (1986). 'Conditionals.' In *Proceedings from the 22nd Regional Meeting of the Chicago Linguistic Society*, 1–15. Chicago: University of Chicago. URL http://semanticsarchive.net/Archive/ThkMjYxN/Conditionals.pdf.

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

Kratzer, A (1990). 'How Specific is a Fact?' In *Proceedings of the 1990 Conference on Theories of Partial In- formation*. Center for Cognitive Science, University of Texas at Austin.

Kratzer, A (1991). 'Modality.' In A von Stechow & D Wunderlich (eds.), *Semantics: An International Handbook of Contemporary Research*, 639–650. Berlin: De Gruyter Mouton.

Kratzer, A (2002). 'Facts: Particulars or Information Units?' *Linguistics and Philosophy*, **25(5–6)**: 655–670.

Kratzer, A (2012). *Modals and Conditionals: New and Revised Perspectives*. New York: Oxford University Press.

Kray, LJ, George, LG, Liljenquist, KA, Galinsky, AD, Tetlock, PE & Roese, NJ (2010). 'From what might have been to what must have been: Counterfactual thinking creates meaning.' *Journal of personality and social psychology*, **98(1)**: 106–118.

Kripke, SA (1963). 'Semantical Analysis of Modal Logic I: Normal Modal Propositional Calculi.' *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, **9**: 67–96.

Kvart, I (1986). *A Theory of Counterfactuals*. Indianapolis, IN: Hackett.

Kvart, I (1992). 'Counterfactuals.' *Erkenntnis*, **36(2)**: 139–179. URL http://dx.doi.org/10.1007/BF00217472.

Lange, M (1999). 'Laws, Counterfactuals, Stability, and Degrees of Likelihood.' *Philosophy of Science*, **66(2)**: 243–267.

Lange, M (2000). *Natural laws in scientific practice*. New York: Oxford University Press.

Lange, M (2009). *Laws and lawmakers: science, metaphysics, and the laws of nature*. Oxford: Oxford University Press.

Leitgeb, H (2012a). 'A Probabilistic Semantics for Counterfactuals: Part A.' *The Review of Symbolic Logic*, **5(1)**: 26–84. URL http://journals.cambridge.org/article_S1755020311000153.

Leitgeb, H (2012b). 'A Probabilistic Semantics for Counterfactuals: Part B.' *The Review of Symbolic Logic*, **5(1)**: 85–121. URL http://journals.cambridge.org/article_S1755020311000165.

Levi, I (1988). 'The Iteration of Conditionals and the Ramsey Test.' *Synthése*, **76(1)**: 49–81.

Lewis, CI (1912). 'Implication and the Algebra of Logic.' *Mind*, **21(84)**: 522–531. URL http://www.jstor.org/stable/2249157.

Lewis, CI (1914). 'The Calculus of Strict Implication.' *Mind*, **23(90)**: 240–247. URL http://www.jstor.org/stable/2248841.

Lewis, D (1981). 'Ordering Semantics and Premise Semantics for Counterfactuals.' *Journal of Philosophical Logic*, **10(2)**: pp. 217–234. URL http://www.jstor.org/stable/30227190.

Lewis, DK (1973a). 'Causation.' *Journal of Philosophy*, **70(17)**: 556–567.

Lewis, DK (1973b). *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press.

Lewis, DK (1973c). 'Counterfactuals and Comparative Possibility.' *Journal of Philosophical Logic*, **2(4)**: 418–446.

Lewis, DK (1979). 'Counterfactual Dependence and Time's Arrow.' *Noûs*, **13**: 455–476.

Lewis, KS (2016). 'Elusive Counterfactuals.' *Noûs*, **50(2)**: 286–313. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/nous.12085.

Lewis, KS (2017a). 'Counterfactual Discourse in Context.' *Noûs*.

Lewis, KS (2017b). 'Counterfactuals and Knowledge.' In JJ Ichikawa (ed.), *The Routledge Handbook of Epistemic Contextualism*, 411–424. New York: Routledge.

Loewer, B (1976). 'Counterfactuals with Disjunctive Antecedents.' *Journal of Philosophy*, **73(16)**: 531–537.

Loewer, B (1983). 'Information and belief.' *Behavioral and Brain Sciences*, **6(1)**: 75—76.

Loewer, B (2007). 'Counterfactuals and the Second Law.' In H Price & R Corry (eds.), *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*, 293–326. New York: Oxford University Press.

Lowe, EJ (1983). 'A simplification of the logic of conditionals.' *Notre Dame Journal of Formal Logic*, **24(3)**: 357–366. URL https://doi.org/10.1305/ndjfl/1093870380.

Lowe, EJ (1990). 'Conditionals, context, and transitivity.' *Analysis*, **50(2)**: 80–87. URL http://dx.doi.org/10.1093/analys/50.2.80.

Lucas, CG & Kemp, C (2015). 'An Improved Probabilistic Account of Counterfactual Reasoning.' *Psychological Review*, **122(4)**: 700–734.

Lycan, WG (2001). *Real Conditionals*. Oxford: Oxford University Press.

Lyons, J (1977). *Semantics*, vol. 2. Cambridge, England: Cambridge University Press.

Mackie, JL (1974). *The Cement of the Universe: A Study in Causation*. Oxford: Oxford University Press.

Marr, D (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.

Maudlin, T (2007). *Metaphysics Within Physics*. New York: Oxford University Press.

McKay, TJ & van Inwagen, P (1977). 'Counterfactuals with Disjunctive Antecedents.' *Philosophical Studies*, **31**: 353–356.

Morreau, M (2009). 'The Hypothetical Syllogism.' *Journal of Philosophical Logic*, **38(4)**: 447–464. URL https://doi.org/10.1007/s10992-008-9098-y.

Morreau, M (2010). 'It Simply Does Not Add Up: trouble with overall similarity.' *The Journal of Philosophy*, **107(9)**: 469–490. URL http://www.jstor.org/stable/29778047.

Moss, S (2012). 'On the Pragmatics of Counterfactuals.' *Noûs*, **46(3)**: 561–586. URL http://dx.doi.org/10.1111/j.1468-0068.2010.00798.x.

Nelson, EJ (1933). 'On Three Logical Principles in Intension.' *The Monist*, **43(2)**: 268–284.

Nozick, R (1969). 'Newcomb's Problem and Two Principles of Choice.' In N Rescher (ed.), *Essays in Honor of Carl G. Hempel*, 111–5. Dordrecht: D. Reidel.

Nute, D (1975a). 'Counterfactuals.' *Notre Dame J. Formal Logic*, **16(4)**: 476–482. URL http://dx.doi.org/10.1305/ndjfl/1093891882.

Nute, D (1975b). 'Counterfactuals and the Similarity of Worlds.' *The Journal of Philosophy*, **72(21)**: 773–8.

Nute, D (1980a). 'Conversational Scorekeeping and Conditionals.' *Journal of Philosophical Logic*, **9(2)**: pp. 153–166. URL http://www.jstor.org/stable/30226202.

Nute, D (ed.) (1980b). *Topics in Conditional Logic*. Dordrecht: Reidel.

Palmer, FR (1986). *Mood and Modality*. Cambridge, England: Cambridge University Press.

Parisien, C & Thagard, P (2008). 'Robosemantics: How Stanley the Volkswagen Represents the World.' *Minds & Machines*, **18(2)**: 169–178.

Pearl, J (1995). 'Causation, Action, and Counterfactuals.' In A Gammerman (ed.), *Computational Learning and Probabilistic Reasoning*, 235–255. New York: John Wiley and Sons.

Pearl, J (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Pearl, J (2002). 'Reasoning with Cause and Effect.' *AI Magazine*, **23(1)**: 95–112. URL http://ftp.cs.ucla.edu/pub/stat_ser/r265-ai-mag.pdf.

Pearl, J (2009). *Causality: Models, Reasoning, and Inference*. 2nd edn. Cambridge, England: Cambridge University Press.

Pearl, J (2013). 'Structural Counterfactuals: A Brief Introduction.' *Cognitive Science*, **37(6)**: 977–985. URL http://dx.doi.org/10.1111/cogs.12065.

Peirce, CS (1896). ‘The Regenerated Logic.’ *The Monist*, **7(1)**: pp. 19–40. URL http://www.jstor.org/stable/27897385.

Pendlebury, M (1989). ‘The Projection Strategy and the Truth Conditions of Conditional Statements.’ *Mind*, **98(390)**: pp. 179–205. URL http://www.jstor.org/stable/2255125.

Pereboom, D (2014). *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.

Pollock, JL (1976). *Subjunctive Reasoning*. Dordrecht: D. Reidel Publishing Co. URL http://oscarhome.soc-sci.arizona.edu/ftp/PAPERS/Pollock_Subjunctive_Reasoning.pdf.

Pollock, JL (1981). ‘A Refined Theory of Counterfactuals.’ *Journal of Philosophical Logic*, **10(2)**: pp. 239–266. URL http://www.jstor.org/stable/30227192.

Quine, WVO (1960). *Word and Object*. Cambridge, MA: MIT Press.

Quine, WVO (1982). *Methods of Logic*. 4th edn. Cambridge, MA: Harvard University Press.

Rips, LJ (2010). ‘Two Causal Theories of Counterfactual Conditionals.’ *Cognitive Science*, **34(2)**: 175–221. URL http://dx.doi.org/10.1111/j.1551-6709.2009.01080.x.

Rips, LJ & Edwards, BJ (2013). ‘Inference and Explanation in Counterfactual Reasoning.’ *Cognitive Science*, **37(6)**: 1107–1135. URL http://dx.doi.org/10.1111/cogs.12024.

Ryle, G (1949). *The Concept of Mind*. London: Hutchinson.

Sanford, DH (1989). *If P then Q: Conditionals and the Foundations of Reasoning*. London: Routledge.

Santorio, P (2014). ‘Filtering Semantics for Counterfactuals: Bridging Causal Models and Premise Semantics.’ *Proceedings of Semantics and Linguistic Theory (SALT) 24*, 494–513. URL http://dx.doi.org/10.3765/salt.v24i0.2430.

Schaffer, J (2016). ‘Grounding in the image of causation.’ *Philosophical Studies*, **173(1)**: 49–100. URL https://doi.org/10.1007/s11098-014-0438-1.

Schulz, K (2007). *Minimal Models in Semantics and Pragmatics: Free choice, Exhaustivity, and Conditionals*. Ph.D. thesis, University of Amsterdam: Institute for Logic, Language and Computation, Amsterdam. URL http://www.illc.uva.nl/Publications/Dissertations/DS-2007-04.text.pdf.

Schulz, K (2011). ‘If you’d wiggled A, then B would’ve changed.’ *Synthese*, **179**: 239–251. URL http://dx.doi.org/10.1007/s11229-010-9780-9.

Schulz, K (2014). 'Fake Tense in conditional sentences: a modal approach.' *Natural Language Semantics*, **22(2)**: 117–144. URL http://dx.doi.org/10.1007/s11050-013-9102-0.

Seto, E, Hicks, JA, Davis, WE & Smallman, R (2015). 'Free Will, Counterfactual Reflection, and the Meaningfulness of Life Events.' *Social Psychological and Personality Science*, **6(3)**: 243–250. URL https://doi.org/10.1177/1948550614559603.

Sider, T (2010). *Logic for Philosophy*. New York: Oxford University Press.

Skyrms, B (1981). 'The Prior Propensity Account of Subjunctive Conditionals.' In W Harper, R Stalnaker & G Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time*, 259–265. Dordrecht: Reidel.

Sloman, S (2005). *Causal Models: How People Think About the World and Its Alternatives*. New York: OUP.

Sloman, SA & Lagnado, DA (2005). 'Do We "do"?' *Cognitive Science*, **29(1)**: 5–39. URL http://dx.doi.org/10.1207/s15516709cog2901_2.

Slote, M (1978). 'Time in Counterfactuals.' *Philosophical Review*, **7(1)**: 3–27.

Smilansky, S (2000). *Free Will and Illusion*. New York: Oxford University Press.

Snider, T & Bjorndahl, A (2015). 'Informative counterfactuals.' *Semantics and Linguistic Theory*, **25**: 1–17. URL http://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/25.01.

Spirtes, P, Glymour, C & Scheines, R (1993). *Causation, Prediction, and Search*. Berlin: Springer-Verlag.

Spirtes, P, Glymour, C & Scheines, R (2000). *Causation, Prediction, and Search*. 2 edn. Cambridge, Massachusetts: The MIT Press.

Sprigge, TL (2006). 'My Philosophy and Some Defence of It.' In P Basile & LB McHenry (eds.), *Consciousness, Reality and Value: Essays in Honour of T. L. S. Sprigge*, 299–321. Heusenstamm: Ontos Verlag.

Sprigge, TLS (1970). *Facts, Worlds and Beliefs*. London: Routledge & K. Paul.

Stalnaker, R (1968). 'A Theory of Conditionals.' In N Rescher (ed.), *Studies in Logical Theory*, 98–112. Oxford: Basil Blackwell.

Stalnaker, R (1972/1981). 'Letter to David Lewis.' In W Harper, R Stalnaker & G Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time*, 151–2. Dordrecht: D. Reidel.

Stalnaker, R (1975). 'Indicative Conditionals.' *Philosophia*, **5**: 269–286. Page references to reprint in Stalnaker (1999).

Stalnaker, RC (1978). 'Assertion.' In P Cole (ed.), *Syntax and Semantics 9: Pragmatics*, 315–332. New York: Academic Press. References to Stalnaker 1999.

Stalnaker, RC (1981). 'A Defense of Conditional Excluded Middle.' In WL Harper, R Stalnaker & G Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time*, 87–104. Dordrecht: D. Reidel Publishing Co.

Stalnaker, RC (1984). *Inquiry*. Cambridge, MA: MIT Press.

Stalnaker, RC (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.

Stalnaker, RC & Thomason, RH (1970). 'A Semantic Analysis of Conditional Logic.' *Theoria*, **36**: 23–42.

Starr, WB (2012). 'The Structure of Possible Worlds.' Talk Delivered at UCLA. URL http://williamstarr.net/research/the_structure_of_possible_worlds.pdf.

Starr, WB (2014). 'A Uniform Theory of Conditionals.' *Journal of Philosophical Logic*, **43(6)**: 1019–1064. URL http://dx.doi.org/10.1007/s10992-013-9300-8.

Stone, M (1997). 'The Anaphoric Parallel between Modality and Tense.' *Tech. Rep. 97–06*, University of Pennsylvania Institute for Research in Cognitive Science, Philadelphia, PA. URL http://www.cs.rutgers.edu/~mdstone/pubs/ircs97-06.pdf.

Swanson, E (2012). 'Conditional Excluded Middle without the Limit Assumption.' *Philosophy and Phenomenological Research*, **85(2)**: 301–321. URL http://dx.doi.org/10.1111/j.1933-1592.2011.00507.x.

Tadeschi, P (1981). 'Some Evidence for a Branching-Futures Semantic Model.' In P Tedeschi & A Zaenen (eds.), *Syntax and Semantics: Tense and Aspect*, vol. 14, 239–69. New York: Academic Press.

Tarski, A (1936). 'Der Wahrheitsbegriff in den formalizierten Sprachen.' *Studia Philosophica*, **1**: 261–405.

Thrun, S, Montemerlo, M, Dahlkamp, H, Stavens, D, Aron, A, Diebel, J, Fong, P, Gale, J, Halpenny, M, Hoffmann, G, Lau, K, Oakley, C, Palatucci, M, Pratt, V, Stang, P, Strohband, S, Dupont, C, Jendrossek, LE, Koelen, C, Markey, C, Rummel, C, van Niekerk, J, Jensen, E, Alessandrini, P, Bradski, G, Davies, B, Ettinger, S, Kaehler, A, Nefian, A & Mahoney, P (2006). 'Stanley: The robot that won the DARPA Grand Challenge.' *Journal of Field Robotics*, **23(9)**: 661–692. URL http://dx.doi.org/10.1002/rob.20147.

Tichý, P (1976). 'A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals.' *Philosophical Studies*, **29**: 271–273.

Todd, W (1964). 'Counterfactual Conditionals and the Presuppositions of Induction.' *Philosophy of Science*, **31(2)**: pp. 101–110. URL http://www.jstor.org/stable/185987.

Van Fraassen, BC (1966). 'Singular Terms, Truth-Value Gaps and Free Logic.' *Journal of Philosophy*, **3**: 481–495.

Veltman, F (1976). 'Prejudices, Presuppositions and the Theory of Counterfactuals.' In J Groenendijck & M Stokhof (eds.), *Amsterdam Papers in Formal Grammar*, Proceedings of the 1st Amsterdam Colloquium, 248–281. University of Amsterdam.

Veltman, F (1985). *Logics for Conditionals*. Ph.D. dissertation, University of Amsterdam, Amsterdam.

Veltman, F (1986). 'Data Semantics and the Pragmatics of Indicative Conditionals.' In EC Traugott, A ter Meulen, JS Reilly & CA Ferguson (eds.), *On Conditionals*. Cambridge, England: Cambridge University Press.

Veltman, F (2005). 'Making Counterfactual Assumptions.' *Journal of Semantics*, **22**: 159–180. URL http://staff.science.uva.nl/~veltman/papers/FVeltman-mca.pdf.

Walters, L (2014). 'Against Hypothetical Syllogism.' *Journal of Philosophical Logic*, **43(5)**: 979–997. URL http://www.jstor.org/stable/24564024.

Walters, L & Williams, JRG (2013). 'An Argument for Conjunction Conditionalization.' *The Review of Symbolic Logic*, **6**: 573–588. URL http://journals.cambridge.org/article_S1755020313000191.

Warmbrōd, K (1981a). 'Counterfactuals and Substitution of Equivalent Antecedents.' *Journal of Philosophical Logic*, **10(2)**: 267–289. URL http://www.jstor.org/stable/30227193.

Warmbrōd, K (1981b). 'An Indexical Theory of Conditionals.' *Dialogue, Canadian Philosophical Review*, **20(4)**: 644–664.

Wasserman, R (2006). 'The Future-Similarity Objection Revisited.' *Synthese*, **150(1)**: 57–67.

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

Willer, M (2015). 'Simplifying Counterfactuals.' In T Brochhagen, F Roelofsen & N Theiler (eds.), *20th Amsterdam Colloquium*, 428–437. Amsterdam: ILLC. URL http://semanticsarchive.net/Archive/mVkOTk2N/AC2015-proceedings.pdf.

Willer, M (2017a). 'Lessons from Sobel Sequences.' *Semantics and Pragmatics*, **10(4)**.

Willer, M (2017b). 'Simplifying with Free Choice.' *Topoi*, 1–14. URL https://doi.org/10.1007/s11245-016-9437-5.

Williams, JRG (2010). 'Defending Conditional Excluded Middle.' *Noûs*, **44(4)**: 650–668. URL http://dx.doi.org/10.1111/j.1468-0068.2010.00766.x.

Williamson, T (2005). 'Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking.' *Proceedings of the Aristotelian Society (Hardback)*, **105(1)**: 1–23. URL http://dx.doi.org/10.1111/j.0066-7373.2004.00100.x.

Williamson, T (2007). *The Philosophy of Philosophy*. Malden, MA: Blackwell.

Woodward, J (2002). 'What is a Mechanism? A Counterfactual Account.' In JA Barrett & JM Alexander (eds.), *PSA'00: Proceedings of the 2000 Biennial Meetings of the Philosophy of Science Association, Part II: Symposium Papers*, S366–S192. Newark, Delaware: Philosophy of Science Association.

Woodward, J (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Zeman, J (1997). 'Peirce and Philo.' In N Houser, D Roberts & JV Evra (eds.), *Studies in the Logic of Charles Sanders Peirce*, 402–417. Indianapolis: Indiana University Press.

Academic Tools

[Auto-inserted by SEP staff]

Other Internet Resources

[Please contact the author with suggestions.]

Related Entries

conditionals: indicative — actualism — impossible worlds — logic: conditionals — logic: modal — modality: epistemology of — modality: varieties of — possible worlds — causation: counterfactual theories of — laws of nature