Causation and probability in indicative and counterfactual conditionals

Daniel Lassiter
Stanford University

Draft as of August 14, 2018.
Comments, criticisms, suggestions, and questions very welcome: danlassiter@stanford.edu.

Abstract This paper provides and motivates a unified compositional semantics for indicative and counterfactual conditionals that have in their consequents probability operators: *probable*, *likely*, *more likely than not*, *50% chance* and so on. The theory combines a Kratzerian syntax for conditionals with semantics based on causal Bayes nets, and develops an algorithm for interpreting probabilistic indicatives and counterfactuals on which the key difference between them involves the choice between two operations on causal models—conditioning and intervention. The interaction between probability operators and the two kinds of conditionals sheds new light on the vexed question of what real-world information is retained in counterfactual scenarios and what is discarded, and helps to explain why indicatives are not sensitive to details of causal structure while counterfactuals are.

The second part of the paper develops a new empirical challenge for the revision procedure that emerges due to the interaction of counterfactuals and probability operators. The straightforward approach to revision developed in the first part is shown to be inadequate for certain scenarios in which real-world information that is discarded in the counterfactual scenario nevertheless informs the values of other variables. I describe two extensions of the semantics in the first part that resolve this puzzle—one based on a more complicated definition of counterfactual probabilities, and one that relies on the related formalism of structural causal models.

Keywords: Indicative conditionals, counterfactual conditionals, epistemic modals, probability operators, causation, causal models, Bayesian networks

1 Conditionals, probability operators, and time

Here is an attractively simple, but flawed, attempt to account for the relationship between two kinds of conditionals in terms of the flow of time. (This is related to, but simplified relative to, ideas discussed by Adams (1975); Slote (1978); Edgington (1995); Khoo (2015) among others.) In many cases, the future indicative (2a) and the retrospective counterfactual (2b) seem to pattern together in terms of truth and/or assertibility.

(1) a. [Spoken on Monday:] If Mary plays in the game Tuesday, we will win.
b. [Mary doesn't play, and we lose. Spoken on Wednesday:] If Mary had played in the game Tuesday, we would have won.

The same holds for overtly probabilistic conditionals:

(2) a. [Spoken on Monday:] If Mary plays on Tuesday, we will probably win.

b. [Mary doesn't play, and we lose. Spoken on Wednesday:] If Mary had played on Tuesday, we probably would have won.

The idea is roughly: for $t_0 < t_1 < t_2$, an indicative uttered at t_0 about an event at t_1 is true just in case the matched counterfactual is, uttered at t_2. This hints at an attractive theoretical reduction, and a useful heuristic for constructing counterfactual scenarios—"rewind" to just before the antecedent time, making necessary changes to render the antecedent true, and consider what happens when the modified history unfolds again. Call this the RRR heuristic—"Rewind, Revise, Re-run".

However, Slote (1978) describes an example due to Sidney Morgenbesser that makes trouble for these claims about (1)-(2) (see especially Barker 1998). This case is known in the literature as "Morgenbesser's coin". Here is a variant:

(3) [A fair coin will be flipped on Tuesday. On Monday you are offered a bet: pay \$50 and win \$100 if the coin comes up heads, but nothing if it's tails.]

a. [Spoken on Monday:] If you bet, you will win.

b. [You don't bet on Monday, and the coin flip on Tuesday comes up heads. Spoken on Wednesday:] If you had bet, you would have won.

(3a) and (3b) seem to differ, but we should be careful not to misdiagnose the difference. The counterfactual (3b), spoken on Wednesday, is plainly true: the coin came up heads, and so if you had bet on heads you would have won. What is the status of (3a)? Well, it is clearly not *assertible* on Monday: there was no reason to favor heads over tails then. But it is plausible that it is *true* on Monday nonetheless, since heads did come up. The speaker made a lucky guess, but lucky guesses can be correct (cf. Edgington 2003: 23; Khoo 2015). This example might still be problematic depending on your position about the truth-aptness of future contingents (e.g., Belnap & Green 1994; MacFarlane 2003). But it does not yet refute the thesis that counterfactuals stand and fall with appropriately matched indicatives.

A variant involving probability operators does refute this thesis, though.

(4) a. [Monday:] If you bet, there's an exactly 50% probability that you'll win.

b. [You don't bet on Monday, and the coin flip on Tuesday comes up heads. Spoken on Wednesday:] If you had bet, there's an (exactly) 50% probability that you'd have won.

(4a), as spoken on Monday, is true. (4b) is false, though: since the coin came up heads, there is a 100% probability on Wednesday that you'd have won if you'd bet.

Morgenbesser's coin also shows that the RRR heuristic does not accurately describe our interpretative procedure for counterfactuals. The problem is that the coin flip is a *random* process, and it happened *after* you declined the bet. Rewind to Monday, Revise so that you take the bet, and Re-run history with this modification. With probability .5, the coin comes up tails on Tuesday and you lose. So the RRR heuristic predicts (4b) to be true, when in fact it is false.[1]

[1] Lewis's (1973) similarity-based account might seem to do better here, since a world in which you bet and the coin flip keeps its actual value (heads) is, in an intuitive sense, more similar to the actual world than one in which you bet and the coin flip has a non-actual value (tails). However, this analysis is not really available to the similarity semantics. Fine

Barker (1998, 1999) gives a diagnosis of the crucial feature of Morgenbesser examples: the temporally later fact that is held fixed—here, that the coin flip came up heads—is **causally independent** of the antecedent. The coin flip happens after your bet, but its outcome does not depend in any way on the bet. The suggestion, as I will develop it, is that the "Rewind, Revise, Re-run" heuristic is *almost* right. When evaluating a counterfactual you should not throw out *all* information about events posterior to the antecedent time. Instead, you selectively throw out information about the outcomes of events that depend causally on the antecedent, and keep information about events that are causally independent of the antecedent. This corresponds to an elaborated heuristic: Rewind to the antecedent time, Revise to make the antecedent true, and selectively Regenerate following events that depend causally on the antecedent.

Kaufmann (2001a,b), Edgington (2003, 2008), and Hiddleston (2005) take up Barker's suggestion and incorporate it into a semantics for bare counterfactuals in various ways. Edgington also discusses, informally, how to accommodate the explicitly probabilistic examples like (4) that Barker focuses on. Focusing on conditionals with overt probability operators, I will argue that these approaches are on the right track for counterfactuals, and that the modified heuristic just described has a precise and illuminating formal implementation in terms of *interventions* on causal Bayes nets ("CBNs"; Meek & Glymour 1994; Pearl 2000). I will also show that probabilistic indicatives behave differently from matched counterfactuals: modifying causal information does not affect indicatives in the same way. To explain this difference, the compositional semantics will interpret probabilistic indicatives and counterfactuals identically modulo the choice between conditioning and intervening. In the last section I show that the revised heuristic is still empirically inadequate due to the special way that probability operators import subjective information into truth-conditional meaning. I describe three candidate solutions: one that defines counterfactual probabilities piecemeal depending on their causal relationship to the antecedent, one that follows Pearl (2000) in utilizing structural Causal models, and one that relies on an additional mechanisms for recalibrating CBNs.

To keep the paper to a manageable length I will neglect one crucial topic: the relationship between probabilistic conditionals and bare conditionals. Somewhat paradoxically, the key analytic tools used here—causal Bayes nets and the restrictor theory of conditionals—both render bare conditionals more complicated than those with overt epistemic operators. A number of strategies are available for extending the account given here to bare conditionals of both types, building on Kratzer 1981a, 1991a; Stalnaker & Jeffrey 1994; Kaufmann 2005, and Pearl 2000. All of these approaches should in principle be compatible with the project pursued here.[2]

(1975) points out that *If Nixon had pressed the button, there would have been a nuclear holocaust* is intuitively true, even though a world in which the button malfunctions and no holocaust occurs is intuitively far more similar to the actual world than one in which the button works normally and a holocaust does occur. Tichý (1976) presents a more complicated example in which Lewis' semantics seems to predict a counterfactual true, based on similarity to facts in the actual world, when it is clearly not true. (The theory advocated in this paper deals with both of these cases easily, by the way.) In response, Lewis (1979) argues that the relevant kind of similarity differs from the intuitive notion, and that similarity in terms of contingent facts after the antecedent time is not relevant. If so, Lewis' theory does not predict the right result in the case of Morgenbesser's coin, since the outcome of the coin toss is a contingent fact that occurs after the antecedent time, and so it should not influence the similarity of worlds. See Edgington 2003, 2011 for careful discussion of this point.

[2] While similarity-based and premise-semantic accounts have long been dominant in philosophy and linguistics, many

2 Causal relations are crucial—but only for counterfactuals

Causal relations are directly relevant to the interpretation of probabilistic counterfactuals, but not to probabilistic indicatives. In this section I will give empirical evidence for this claim and begin to flesh out the "Rewind, Revise, selectively Regenerate" heuristic described in §1 more precisely. §3 implements the revised heuristic formally, shows how it interacts with causal and probabilistic information to generate counterfactual probabilities, and explains why indicatives differ.

Consider the following scenario, which verifies the counterfactuals in (5). (This morbid tale is modified from Edgington 2003.)

> **Version 1**: On the way to the airport to fly to Paris, Fran's car gets a flat tire. She misses her flight. During the flight, the pilot has a heart attack, the plane crashes, and 70% of the passengers are killed.

(5) a. If Fran had made her flight, it is likely/probable that she would have died.

b. If Fran had made her flight, there is a 70% chance she would have died.

These are Morgenbesser counterfactuals: their truth depends on holding fixed two contingent propositions, **heart-attack** and **crash**, that occur after the time of the antecedent **made-the-flight**. Since **heart-attack** and **crash** are causally independent of whether Fran made her flight, the revised, causally sensitive heuristic—"Rewind, Revise, selectively Regenerate"—holds them fixed in the counterfactual scenario.

Consider now a slightly enriched version of the story:

> **Version 2**: Same as Version 1, except that Fran is a highly skilled pilot who could easily land a passenger jet safely.

In this context, the probabilistic counterfactuals in (5) seem to be false. The revised heuristic gives a sense of how this is possible. Once we add the information that Fran is a pilot, the crash is no longer causally independent of whether she is aboard. Instead, there is a causal connection between her presence and the crash—since she could have landed the plane safely, whether the crash occurs depends on both **heart-attack** and **made-the-flight**. The revised heuristic therefore instructs us to ignore the contingent information that the crash happened.

As a step toward formalizing this reasoning, consider the simplified causal models of these scenarios in Fig. 1. Nodes pick out questions of interest (partitions on W, or "variables" in statistical jargon). Arrows indicate direct causal connections: $Q_1 \rightarrow Q_2$ indicates that the answer to Q_1 is an

theorists have discussed the idea that counterfactuals depend on causation rather than vice versa—too many to undertake a detailed comparison with all relevant previous work in this space, unfortunately. Here is a partial list of work not cited in the main text, focusing just on philosophy and linguistics: Jackson 1977; Kvart 1986, 1992; Schaffer 2004; Schulz 2007, 2011; Briggs 2012; Kaufmann 2013; Stalnaker 2015; Zhao 2015; Santorio 2016. If we included psychology (both cognitive and social), computer science, and statistics the list would be much longer. I have been inspired by many of these sources, and have focused on the ones discussed in the main text primarily because they bring out some of the key issues that I am concerned with here—causal relevance and the interaction with the language of probability. In particular, the large majority of applications of causal models in formal semantics have focused on deterministic models, which are logically simpler but insufficient to account for the phenomena considered here.

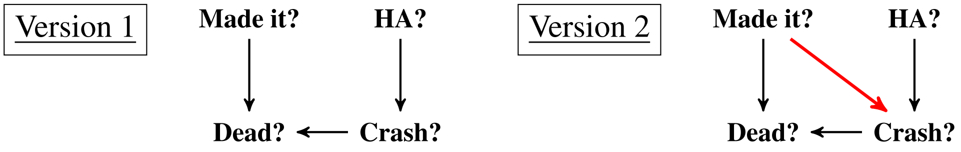


Figure 1 Causal structure of the two stories, with changes in version 2 in red.

input to the function that determines the probability distribution on answers to Q_2. If there is a path $Q_1 \rightarrow \ldots \rightarrow Q_n$, we say that Q_1 is *upstream*/an *ancestor* of Q_n, and Q_n is *downstream* of Q_1.

In the causal model of version 1 on the left side of Figure 1, there is no path from **Made-it?** to either **HA?** ("Did the heart attack occur?") or **Crash?**. Instead, **Made-it?** is relevant only to the question **Dead?**: Fran presence and whether the crash occurs jointly determine whether she dies. This models the intuition that the heart attack and the plane crash are both causally independent of Fran's presence in the first version. The only variable that depends on whether she is aboard is whether she dies. In the second version the question of whether Fran makes the flight is causally relevant to the question of whether the plane crash occurs. Modeling this dependence requires us to add an arrow from **Made it?** to **Crash?**.

We can now state the revised heuristic a bit more precisely:

> When evaluating a counterfactual, (i) identify the question Q that the antecedent answers, (ii) discard the current answer to Q, (iii) add the information that the antecedent is true, and (iv) regenerate portions of the model that are downstream of Q, ignoring the actual answers to questions that are causally downstream of the antecedent. Do not change anything that is not causally downstream of the antecedent.

Note that the heuristic no longer refers to time. While the causal modeling approach does not require us to take a strong stance on the issue, it suggests the tantalizing possibility that the apparent relationship between counterfactuals and time may be a mere side effect of the fact that causal relations unfold over time—i.e., causes invariably precede their effects. The phenomena discussed in this paper seem to be consistent with this strong stance.[3]

This heuristic handles the probabilistic Morgenbesser example (4) straightforwardly. The causal structure is just **bet?** → **win?** ← **heads?**, with the question of whether you won a joint effect of two causally unrelated events—whether you bet, and how the coin comes up. When revising to make **bet?** true, we throw out the downstream fact that **win?** is false, but keep the causally independent fact that **Heads?** is true. In the revised model, you bet and the coin comes up heads. *There is an exactly 50% chance that you win* is false in this counterfactual scenario: the true probability is 1. (See section 3.4 below for a precise compositional derivation of this result.)

We now have an explanation of why, in the flight scenario, intuitions about the sentences in (5)

[3] Step (i) hides a complexity: how do we proceed with complex antecedents, when there may not be a question that corresponds to the antecedent? I will not deal with this issue here, but see Briggs 2012; Ciardelli, Zhang & Champollion 2018; Lassiter 2017a for some starting points.

can be reversed by the addition of a causal connection between Fran's presence and the crash. In Version 1 (Figure 1, left), **Crash?** is causally independent of **Made it?**. When we revise to make **Made it?** true, we throw out the downstream observation that Fran is not dead (**Dead?** = F), but we hold fixed that the crash happened, since this question is causally independent of the antecedent. In a revised model with these features—there is a crash, and Fran is on board—we can make a prediction about the value of **Dead?** based on background knowledge about the causal structure of the scenario. The results are pretty grim: since 70% of the passengers died in the crash, presumably there is a 70% chance that Fran dies. This part of the reasoning is still informal, but we will make it precise in the next section.

The revised heuristic tells us to proceed differently in version 2, where Fran's presence on the flight is causally relevant to the crash (Figure 1, right). In this case, we throw out not only the fact that Fran is not dead, but also the fact that the crash happened, since both are downstream of the antecedent question **Made-it?**. The only question in the model whose answer we keep fixed is **HA?**—we hold fixed the causally independent fact that the pilot had a heart attack. In the revised scenario, Fran is aboard, the pilot is incapacitated, and we must use background causal knowledge to determine whether the crash happens and whether Fran dies. Based on the informal description given above, we may expect that Fran will save the day, the crash will not happen, and she will not die. (Again, we will see a formal model that derives this prediction in full in the next section.)

Manipulations of causal relevance lead to an interesting reversal of intuitions about the probabilistic counterfactuals in (5)—and we have the beginnings of a theory of counterfactual interpretation that makes sense of this reversal. What is the status of indicatives? Consider a variant of the original story, revised slightly to satisfy the felicity requirements of the indicative (i.e., that its antecedent's truth-value should not be known).

> **Version 1—unknown**: Fran was supposed to fly to Paris, but her car got a flat tire on the way to the airport. We don't know if she made the flight or not. We do know the pilot had a heart attack, the plane crashed, and 70% of the passengers were killed.

Against this background, consider the probabilistic indicatives corresponding to (5).

(6) a. If Fran made her flight, it is likely/probable that she died.
b. If Fran made her flight, there is a 70% chance that she died.

The sentences in (6) are, regrettably, true; we can only hope that Fran is not skilled at changing tires. Do intuitions reverse when we learn that Fran is a pilot?

> **Version 2—unknown**: Same as Version 1—unknown, except that Fran is a highly skilled pilot who could easily land a passenger jet.

No: once we know that the plane crashed, changing the story so that Fran is a potential hero does not make a difference to the probabilistic indicative. The fact of the crash is held fixed, even though it is causally downstream of the antecedent.

We will now formalize the interaction between conditional type and causal structure, with special attention to deriving the needed indicative and counterfactual probabilities while keeping the interpretation procedure maximally uniform.

3 Indicative vs. counterfactual as conditioning vs. intervening

3.1 Causal Bayes nets

Causal Bayes nets (CBNs) can be defined using familiar concepts from intensional and degree semantics.

(7) A causal Bayes net B is given by $\langle W, \mathcal{Q}, \mathcal{A}, \mathcal{C} \rangle$, where

a. W is a set of possible worlds (the “sample space”).

b. $\mathcal{Q}$ is a set of questions (“variables”), where each $Q \in \mathcal{Q}$ is a partition on W.

c. $\mathcal{A}$ is an acyclic binary relation on $\mathcal{Q}$ (the “arrows”). $\langle Q_i, Q_j \rangle \in \mathcal{A}$ means that Q_i is one of Q_j’s parents, and has an immediate causal influence on Q_j.[4]

d. $\mathcal{C}$ is a set of conditional probability tables that determine a unique conditional probability distribution $P(Q_i \mid Parents(Q_i))$ for each question $Q_i \in \mathcal{Q}$.

I assume the standard definitions of (finitely additive) probability: P’s domain is an algebra of subsets of W, $P(W) = 1$, and $P(X \cup Y) = P(X) + P(Y)$ whenever $X \cap Y = \varnothing$. The conditional probability $P(A \mid B)$ is $P(A \cap B)/P(B)$ whenever $P(B) \neq 0$, undefined otherwise. In addition, the probability measure P associated with Bayes net B is required to obey the *Markov condition*: each variable is probabilistically independent of its non-descendents, given the values of its parents.

A CBN, as defined here, does not generally determine a unique probability measure P. This is because the set of questions $\mathcal{Q}$ may provide only a very coarse grain on the set of possible worlds. However, we can simplify by assuming that there is exactly one possible world for each cell in the maximally fine-grained question given by the intersection of the questions in $\mathcal{Q}$. For example, in the causal models in Figure 1 there were four questions, each with two answers (true/false). The simplification amounts to the assumption that there are at most 2^4 possible worlds—one for each combination of answers to the four questions. When this assumption is satisfied, each causal Bayes net B will determine a prior probability measure P_B.

For example, here are two CBNs that result from enriching the models in Figure 1 with conditional probability tables. (The conditional probability distributions used in Fig. 1 are meant to represent sensible assumptions, but are fairly arbitrary in the details.) Note that the distribution on **C?** in Version 1 depends directly only on **HA?**, and is (by the Markov condition) independent given a value of **C?** of any other variable except its daughter **D?**. In the model of Version 2, the addition of a causal relation between **M?** and **C?** requires a more complex model of the distribution on **C?**. In this case, the probability of **C?** depends on both **M?** and **HA?**.

These models allow us to answer many useful queries. For example, in Version 1 the probability of death, given that Fran made it and the pilot had a heart attack, is

$$\begin{aligned} P(\mathbf{D} \mid \mathbf{M} \cap \mathbf{HA}) &= P(\mathbf{D} \mid \mathbf{M} \cap \mathbf{C}) \times P(\mathbf{C} \mid \mathbf{HA}) + P(\mathbf{D} \mid \mathbf{M} \cap \overline{\mathbf{C}}) \times P(\overline{\mathbf{C}} \mid \mathbf{HA}) \\ &= .7 \times 1 + .001 \times 0 = .7 \end{aligned}$$

[4] Note that this relation is automatically acyclic if we assume that causes precede effects and that temporal precedence is acyclic. Santorio (2016) considers a case which might motivate considering causal models with cyclic causal paths. If so, we can no longer think of the model as describing token causation, as I have been assuming.

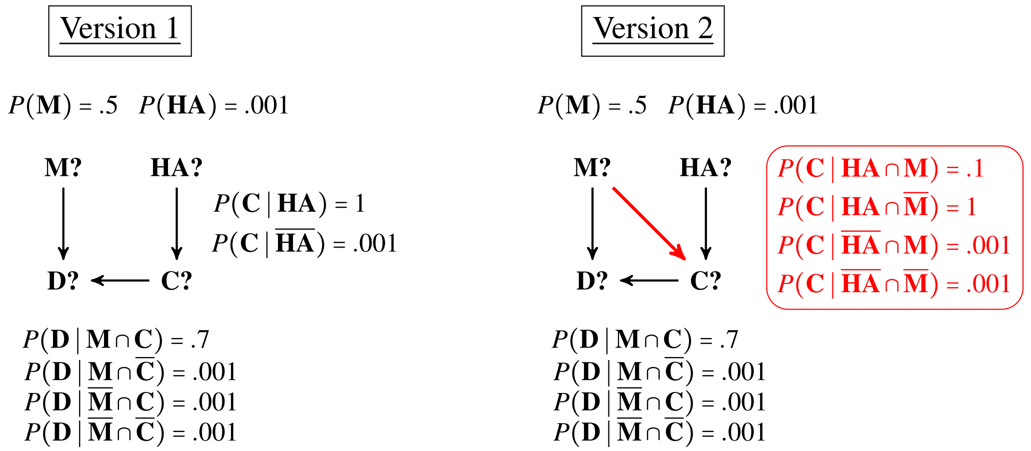

Version 1
$P(\mathbf{M}) = .5$ $P(\mathbf{HA}) = .001$
M?
HA?
D?
C?
$P(\mathbf{C} \mid \mathbf{HA}) = 1$
$P(\mathbf{C} \mid \overline{\mathbf{HA}}) = .001$
$P(\mathbf{D} \mid \mathbf{M} \cap \mathbf{C}) = .7$
$P(\mathbf{D} \mid \mathbf{M} \cap \overline{\mathbf{C}}) = .001$
$P(\mathbf{D} \mid \overline{\mathbf{M}} \cap \mathbf{C}) = .001$
$P(\mathbf{D} \mid \overline{\mathbf{M}} \cap \overline{\mathbf{C}}) = .001$
Version 2
$P(\mathbf{M}) = .5$ $P(\mathbf{HA}) = .001$
M?
HA?
D?
C?
$P(\mathbf{C} \mid \mathbf{HA} \cap \mathbf{M}) = .1$
$P(\mathbf{C} \mid \mathbf{HA} \cap \overline{\mathbf{M}}) = 1$
$P(\mathbf{C} \mid \overline{\mathbf{HA}} \cap \mathbf{M}) = .001$
$P(\mathbf{C} \mid \overline{\mathbf{HA}} \cap \overline{\mathbf{M}}) = .001$
$P(\mathbf{D} \mid \mathbf{M} \cap \mathbf{C}) = .7$
$P(\mathbf{D} \mid \mathbf{M} \cap \overline{\mathbf{C}}) = .001$
$P(\mathbf{D} \mid \overline{\mathbf{M}} \cap \mathbf{C}) = .001$
$P(\mathbf{D} \mid \overline{\mathbf{M}} \cap \overline{\mathbf{C}}) = .001$

Figure 2 Causal models including illustrative conditional probability tables specifying the distribution on each question conditional on its parents.

The first part of the equation—$P(\mathbf{D} \mid \mathbf{M} \cap \mathbf{HA}) = P(\mathbf{D} \mid \mathbf{M} \cap \mathbf{C}) \times P(\mathbf{C} \mid \mathbf{HA})$—holds because the conditional independencies encoded in the graph allow you to decompose complex and long-distance queries into chains of simpler, more local queries. The rest is just look-up of values in the conditional probability tables. In the second model this same probability would decompose differently, due to the difference in causal structure: **C** depends on both **M** and **HA**.

$$\begin{aligned} P(\mathbf{D} \mid \mathbf{M} \cap \mathbf{HA}) &= P(\mathbf{D} \mid \mathbf{M} \cap \mathbf{C}) \times P(\mathbf{C} \mid \mathbf{M} \cap \mathbf{HA}) + P(\mathbf{D} \mid \mathbf{M} \cap \overline{\mathbf{C}}) \times P(\overline{\mathbf{C}} \mid \mathbf{M} \cap \mathbf{HA}) \\ &= .7 \times .1 + .001 \times .9 = .0709. \end{aligned}$$

The substantial difference between these conditional probabilities makes sense. In version 1, conditioning on Fran's presence or absence has no effect on the probability of a crash. In version 2, $P(\mathbf{C} \mid \mathbf{HA} \cap \mathbf{M}) \ll P(\mathbf{C} \mid \mathbf{HA} \cap \overline{\mathbf{M}})$: Fran's presence lowers the probability of crash in case of heart attack from 1 to .1. So, the probability of death **D**—which is an effect of the crash—is also much reduced.

3.2 Overtly probabilistic language

I will use a variant (and in some ways simplified, in others more complex) of a probabilistic domain semantics for *likely*, *probable/probably*, and *chance* (Yalcin 2007, 2010).[5] We relativize interpretation to a world w and a Bayes net B. Simplifying as discussed above, I assume that causal Bayes net B determines a unique probability measure P_B. In addition, we relativize to a set of observations $\mathcal{O}$. $\mathcal{O}$ may, but need not, be interpreted as the total information of some agent. (For instance, the background information in the first version of the story would be represented by $\mathcal{O} = \{\textbf{Made-it?} = F, \textbf{Crash?} = T, \textbf{Dead} = F\}$.) The inclusion of $\mathcal{O}$ is crucial because judgments

[5] See also Yalcin 2012; Swanson 2006, 2011, 2015; Lassiter 2010, 2015, 2017b; Moss 2015.

about whether ϕ is likely should not depend only on the prior information about ϕ's likelihood that a CBN encodes—nor should it depend on the full totality of facts about the world, which would be appropriate only when we are modeling objective probabilities or the beliefs of an omniscient agent.

(8) $[\![\text{likely}]\!]^{w,B,\mathcal{O}} = \lambda d_d \lambda q_{\langle s,t \rangle} \, . \, P_B(q \mid \mathcal{O}) > d$

Let's assume that the meaning of *likely* in the positive form is fixed to a threshold of .5 in every context. (This is not quite right, but it is close enough for present purposes.)

(9) $[\![(\mathit{pos}) \text{ likely}]\!]^{w,B,\mathcal{O}} = \lambda q_{\langle s,t \rangle} \, . \, P_B(q \mid \mathcal{O}) > .5$

The effect is that, for any ϕ, $[\![\phi \text{ is likely}]\!]^{w,B,\mathcal{O}} = 1$ iff $P_B(\phi \mid \mathcal{O}) > .5$. I will assume that *It is probable that* ϕ and *Probably* ϕ are equivalent to *It is likely that* ϕ.

When a percentage expression binds *likely*'s degree argument, the result denotes a probability condition that depends on the percentage expression. For instance:

(10) $[\![\text{exactly } 70\% \text{ likely}]\!]^{w,B,\mathcal{O}} = \lambda q_{\langle s,t \rangle} \, . \, P_B(q \mid \mathcal{O}) = 0.7$

As for *chance*, I will sidestep the interesting compositional puzzles presented by *There is a (n%) chance/probability that ...*, assuming simply that both are equivalent to *It is n% likely that* ϕ. I will also ignore important questions about whether there are different kinds of probability involved in the interpretation of these expressions.[6] Note, however, that it is possible to vary the interpretation between objective and subjective probability by varying the content of $\mathcal{O}$. To model "the subjective probability of ϕ is 0.7", we fill this parameter in with the observations that (say) some relevant individual or group has made. To model "the objective probability of ϕ at time t is 0.7", we use all of the actual values of variables whose answer has been determined at or before t. (With the semantics proposed below, the same trick can be used to model subjective vs. objective readings of indicative and counterfactual conditionals, as well as assessor-sensitivity if desired.)

For the moment, I will not commit to any claims about whether the meanings of other epistemic expressions also invoke probability.

3.3 Probabilistic indicatives: Restriction as conditioning

I assume a restrictor syntax (Kratzer 1991a) for both indicative and counterfactual conditionals. In this approach the antecedent is interpreted as modifying the value of a contextual parameter temporarily, for the purpose of evaluating the consequent. If the parameter modified affects the interpretation of an operator *Op* in the consequent, the interpretation of the consequent will be affected as a result. Since the key examples treated here have overt epistemic operators in the consequent, I will not take sides in the debate about how to extend this analysis to bare conditionals, where there is no overt operator for the antecedent to influence.

For probabilistic indicatives, the target interpretation is one in which the probability measure referred to by the operator in the consequent is conditioned on the information in the antecedent

[6] For example, subjective vs. objective: see Ülkümen, Fox & Malle 2015; Lassiter 2018 for evidence that this distinction is relevant to the semantics of English probability expressions.

(Yalcin 2007, 2012). This means, roughly, treating the antecedent as a "virtual observation" for the purpose of evaluating the consequent. We can implement this analysis very simply by adding the antecedent to the set of observations that the probability measure in the consequent is then conditioned on.

(11) $[\![\text{If } \phi,\ Op\ \psi]\!]^{w,B,\mathcal{O}} = [\![Op\ \psi]\!]^{w,B,\mathcal{O}^+}$, where $\mathcal{O}^+ = \mathcal{O} \cup \{[\![\phi]\!]^{w,B,\mathcal{O}}\}$.

Since the probability operators defined above are always conditioned on $\mathcal{O}$, the effect of (11) is to add the interpretation of the antecedent as an additional condition.

For example, in the indicative-friendly **unknown** variants of the flight story, we do not know whether Fran made her flight, or whether she has died. So the observations are $\mathcal{O} = \{\textbf{heart-attack}, \textbf{crash}\}$. (11) tells us to analyze (6a) as follows:

(12) $[\![\text{If Fran made her flight, it is likely that she died}]\!]^{w,B,\mathcal{O}=\{\textbf{HA},\textbf{C}\}}$

$= [\![\text{it is likely that she died}]\!]^{w,B,\mathcal{O}=\{\textbf{HA},\textbf{C},\textbf{M}\}}$

$=$ True iff $P_B(\textbf{D} \mid \{\textbf{HA}, \textbf{C}, \textbf{M}\}) > 0.5$ (by (21))

$=$ True iff $P_B(\textbf{D} \mid \textbf{M} \cap \textbf{C}) > 0.5$ (independencies in B, Fig. 2)

$=$ True iff $.7 > .5$ (look-up in B, Fig. 2)

$=$ True

Note that it does not matter to this reasoning whether there is a causal link between **made-it?** and **crash?**. This is because the definition in (11) simply adds a proposition to $\mathcal{O}$, without regard for how this proposition is connected causally to others. This feature explains why manipulations of causal structure, holding other relevant facts about probabilities fixed, do not matter for indicatives.[7]

3.4 Probabilistic counterfactuals: Restriction as intervention

The proposed interpretation of counterfactuals is minimally different from that of the matched indicatives. The gross syntax of conditional sentences is the same, except that the morphological differences signal different ways to use the antecedent to modify the parameters of evaluation for the purpose of evaluating the consequent.

In the last section I proposed that indicative morphology in the antecedent is associated with conditioning—addition of the antecedent to the set of observations $\mathcal{O}$, with no change to the structure of the background causal model B. Subjunctive morphology, in contrast, is associated with the more complex operation of **intervention** (Meek & Glymour 1994; Pearl 2000; Pearl & Mackenzie 2018), which makes separate modifications to the B and $\mathcal{O}$ parameters. My treatment is in the spirit of these previous accounts, but differs in that the implementation is tailored to make the compositional semantics simple, and is designed to emphasize the independence of the two kinds of

[7] It is possible that $P(\textbf{D} \mid \textbf{M} \cap \textbf{C})$ could be different in the models of the two scenarios, with the result that (12) could vary in truth-value. This would imply that there is some probabilistically relevant difference between the two scenarios that has not been included in the story. (For example, Fran the pilot might know which parts of the plane are less dangerous in a crash.) This does not undermine the thesis pursued here. Estimating parameters and choosing which variables to model might well be affected by the addition of such information, but the fact of the crash is held fixed in this reasoning.

modifications—revisions to $\mathcal{O}$ and to B. Here is a first pass, assuming that ϕ answers a particular question Q_ϕ in the definition of B. Note the close similarity to the rule for interpreting indicatives given in (11).

(13) $[\![\text{If were } \phi, \text{ would } Op\ \psi]\!]^{w,B,\mathcal{O}} = [\![Op\ \psi]\!]^{w,B^*,\mathcal{O}^*}$, where

a. B^* is B possibly with the removal of some causal links (more below).

b. $\mathcal{O}^*$ is $\mathcal{O}$ minus any answer to Q_ϕ or any of its causal descendants, plus ϕ.

The first condition (13a) is used to regulate the availability of backtracking interpretations of counterfactuals. The second condition (13b) implements the "Rewind, Revise, selectively Regenerate" heuristic motivated above. A key advantage of the semantics presented here, to my mind, is to maintain a clean separation between issues around (a) which pieces of factual information are retained and which are ignored when we create counterfactual scenarios, and (b) how and to what extent we can backtrack, inferring from counterfactual antecedents to features of their causal ancestors. We will return to both conditions for further discussion below; but we can already see how this accounts for Edgington's flight examples.

The precise spell-out of (13a) will not matter because the consequent is causally downstream of the antecedent. So let's simplify by assuming that B^* is just B. The target sentence is *If Fran had made her flight, it is likely that she would have died*. For both versions of the story, we are evaluating this sentence against the observational background $\mathcal{O} = \{\overline{\mathbf{M}}, \mathbf{HA}, \mathbf{C}, \overline{\mathbf{D}}\}$. For version 1 (Figure 2, left), where **C?** is causally independent of **M?**, (11) tells us to compute the value of *she probably died* relative to $B^* = B$ and a modified observation set $\mathcal{O}^*$. Here, $\mathcal{O}^*$ is $\mathcal{O}$ minus $\overline{\mathbf{M}}$ (which answers **M?**, the question addressed by the antecedent) and **D** (which is downstream of **M?**), and with the addition of the antecedent **M**. So, $\mathcal{O}^* = \{\mathbf{HA}, \mathbf{C}, \mathbf{M}\}$ for Version 1.

(14) $[\![\text{If Fran had made her flight, it's likely she would have died}]\!]^{w,B,\mathcal{O}=\{\overline{\mathbf{M}},\mathbf{HA},\mathbf{C},\overline{\mathbf{D}}\}}$

$= [\![\text{it's likely she died}]\!]^{w,B,\mathcal{O}^*=\{\mathbf{HA},\mathbf{C},\mathbf{M}\}}$,

We can stop here: the second line is identical to the second line of (12)! So, of course, the counterfactual is true: the counterfactual probability of death given that she made her flight is .7.

The close connection between (12) and (14) illustrates a strong prediction of the present theory about the relationship between probabilistic indicatives and counterfactuals, reminiscent of the flawed efforts discussed in §1 to fix an indicative/counterfactual connection on the basis of temporal information. On the present account, holding causal structure fixed, a probabilistic counterfactual whose antecedent is known to be false should always pattern with the matched indicative, evaluated in a context where the observations are the same except that the value of the antecedent is unknown. In unpublished experimental work I have verified this prediction quantitatively for the items *likely, probably, might, certain*, and *have to*, across a variety of probabilistic contexts.

There are two key differences in Version 2: the conditional probability tables are different (see Figure 2), and step (13b) operates differently. The truth-conditions for Version 2 depend on $\mathcal{O}^* = \mathcal{O} - \{\overline{\mathbf{M}}, \overline{\mathbf{D}}, \mathbf{C}\} \cup \{\mathbf{M}\}$, where the observation that there was a crash (**C**) has also been removed as being causally downstream of the antecedent. This is enough to explain why the addition of a causal link between **M?** and **C?** has the effect that it does:

(15) $[\![\text{If Fran had made her flight, it's likely she would have died}]\!]^{w,B,\mathcal{O}=\{\overline{\mathbf{M}},\mathbf{HA},\mathbf{C},\overline{\mathbf{D}}\}}$

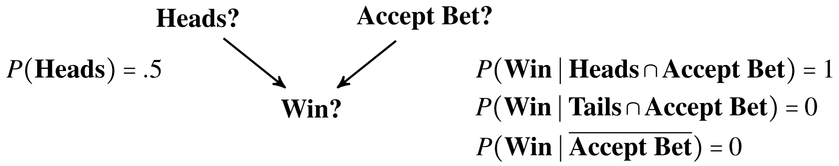

Heads?
Accept Bet?
Win?
$P(\textbf{Heads}) = .5$
$P(\textbf{Win} \mid \textbf{Heads} \cap \textbf{Accept Bet}) = 1$
$P(\textbf{Win} \mid \textbf{Tails} \cap \textbf{Accept Bet}) = 0$
$P(\textbf{Win} \mid \overline{\textbf{Accept Bet}}) = 0$

Figure 3 CBN for the probabilistic Morgenbesser example (4).

$$= [\![\text{it's likely she died}]\!]^{w,B,\mathcal{O}^*=\{\textbf{HA},\textbf{M}\}}$$
$$= \text{True iff } P_B(\textbf{D} \mid \textbf{HA} \cap \textbf{M}) > .5,$$

where the B in question is the CBN in Figure 2, right. As we saw in section 3.1, $P_B(\textbf{D} \mid \textbf{HA} \cap \textbf{M}) > .5$ is just .0709 in Version 2. This is well below .5, and so the sentence is false.

We can also give a precise account of the probabilistic Morgenbesser example (4)—repeated as (16)—that made trouble for the temporally-oriented theory of counterfactual revision described in section 1. Recall that the prospective indicative (16a) was intuitively true, while the matched retrospective counterfactual (16b) was intuitively false.

(16) [A fair coin will be flipped on Tuesday. On Monday you are offered a bet: pay \$50 and win \$100 if the coin comes up heads, but nothing if it's tails.]

a. [Monday:] If you bet, there's an exactly 50% probability that you'll win.

b. [You don't bet on Monday, and the coin flip on Tuesday comes up heads. Spoken on Wednesday:] If you had bet, there's an (exactly) 50% probability that you'd have won.

Figure 3 specifies a CBN for this example. (The distribution on **Accept Bet?** is left unspecified as it does not play a role in the example.)

On Monday, the relevant observation set is $\mathcal{O} = \varnothing$, and the indicative conditional (16a) is true iff $P_B(\textbf{Win} \mid \textbf{Accept Bet}) = .5$—the conditional probability of winning, given that you accept the bet, is exactly .5. This is true, as per intuition.

$$\begin{aligned} P_B(\textbf{Win} \mid \textbf{Accept Bet}) &= P_B(\textbf{Win} \mid \textbf{Heads} \cap \textbf{Accept Bet}) \times P_B(\textbf{Heads}) \\ &\quad + P_B(\textbf{Win} \mid \textbf{Tails} \cap \textbf{Accept Bet}) \times P_B(\textbf{Tails}) \\ &= 1 \times .5 + 0 \times .5 \\ &= .5 \end{aligned}$$

On Wednesday, the relevant observation set is $\mathcal{O} = \{\textbf{Heads}, \overline{\textbf{Accept Bet}}, \overline{\textbf{Win}}\}$. To evaluate a counterfactual with antecedent *If you had bet, ...*, we identify the node that the antecedent answers—**Accept Bet?**—and remove from $\mathcal{O}$ the answer $\overline{\textbf{Accept Bet}}$. In addition, we remove $\overline{\textbf{Win}}$ from $\mathcal{O}$ since it is downstream of **Accept Bet?**. So, $\mathcal{O}^*$ contains only **Heads** and the antecedent **Accept**

Bet, and the counterfactual (16b) is evaluated as

$$\begin{aligned}
&[\![\text{If you had bet, there is an exactly 50\% chance that you would have won}]\!]^{w,B,\mathcal{O}=\{\textbf{Heads},\overline{\textbf{Accept Bet}},\overline{\textbf{Win}}\}}\\
&\qquad = [\![\text{there is an exactly 50\% chance that you would have won}]\!]^{w,B,\mathcal{O}^*=\{\textbf{Heads},\textbf{Accept Bet}\}}\\
&\qquad\quad = \text{True iff } P_B(\textbf{Win} \mid \textbf{Heads} \cap \textbf{Accept Bet}) = .5\\
&\qquad\quad = \text{True iff } 1 = .5\\
&\qquad\quad = \text{False}
\end{aligned}$$

The retrospective counterfactual is correctly ruled false, since the true probability of winning—holding fixed the causally independent fact that the coin come up heads—is 1.

This concludes the analysis of the compositional interaction between probability operators and the two kinds of conditionals. The next sections take up a number of issues with the detailed statement of the interpretation rule for counterfactuals in (13), with particular attention to some new puzzles around counterfactual revision (13b) that become apparent only when we consider the interaction with probability operators in detail.

4 Backtracking

Regarding the underspecified "graph surgery" condition (13a): It has often been thought desirable to rule out interpretations of counterfactuals like (17) which involve reasoning from effects to causes.

(17) [Jar A has 2 blue and 8 red balls. Jar B has 8 red and 2 blue. The ball drawn was blue, but we don't know which jar the ball came from.]
If the ball were red, it would probably have come from Jar B.

In this scenario, the choice of jar determines the probability of each color, via the number of red vs. blue balls. A minimal causal model is **Jar** → **Color**. So, a true reading of (17) involves reasoning from effect **Color** to cause **Jar**, i.e., backtracking.

Lewis (1979), for instance, argues that backtracking readings should be false as a default, and only true on a "special" reading. However, backtracking readings are fully compatible with Lewis' semantic theory: their absence (or "specialness") is enforced by informal conditions on the similarity ordering. Theories of counterfactuals built around causal models have generally taken a stronger stance, ruling out backtracking as part of the definition of intervention (notably Pearl 2000; though see Hiddleston 2005; Lucas & Kemp 2015 for important exceptions.) This idea has an especially simple implementation in Pearl 2000, who defines interventions on a model B relative to a "surgically modified" model $B\star$. $B\star$ is identical to B except that all incoming links are removed from the intervention site (the antecedent, for counterfactuals). In the modified model used to interpret *If the ball were red, ...*, the question "Which color is the ball?" has no parents, and so no causes. As a result, after graph surgery adding information about **Color** cannot influence the now-independent variable **Jar**. For Pearl, an intervention to make the ball red cannot convey any information about which jar the ball came from.

However, there is extensive experimental evidence that English speakers do perceive backtracking interpretations of counterfactuals (Sloman & Lagnado 2005; Rips 2010; Dehghani, Iliev

& Kaufmann 2012; Gerstenberg, Bechlivanidis & Lagnado 2013; Lucas & Kemp 2015). Indeed, in the unpublished experiment mentioned briefly above I also tested the acceptability of (17) and related probabilistic conditionals. Participants did not hesitate to interpret this sentence as being highly acceptable as a description of a visual scene corresponding to the context described—indeed, just as acceptable as the matched indicative *If the ball is red, it probably came from B* in a context where the color is unknown. In the latter case, reasoning from effect to cause is uncontroversially available.[8] These results show that backtracking interpretations of probabilistic counterfactuals are robustly available. So, our theory must make room for these interpretations. At the same time, the theory should not make backtracking obligatory, given the many intuitive (e.g. Lewis 1979; Khoo 2016) and experimental (e.g., Sloman & Lagnado 2005) demonstrations suggesting that non-backtracking interpretations are often more accessible. In the approach pursued here, there is in principle room for both interpretations without affecting the separate issue of causal (in)dependence and its effects on counterfactual revision. In brief, the idea is that "graph surgery" is more flexible than Pearl allows. In the modified graph $B*$ in (13a), the causal links that are removed do not have to be between the intervention site and its immediate parents, but may be further upstream. At the extremes, the modified graph can be the original B (with no links removed) or have all incoming links to the antecedent removed (Pearl's approach). In complex models, there will be many intermediate choices.

The resolution of this ambiguity is somehow context-dependent (Lewis 1979; Kaufmann 2013; Khoo 2016), and it may also depend on the presence of epistemic language in the counterfactual consequent. The invocation of "context" does not, of course, constitute a predictive theory of when and why backtracking readings are available; it only leaves room for them as needed. Hopefully future work will make available a precise statement of when and why backtracking interpretations are available, which can then be integrated into the present theory. (See Arregui 2005; Schulz 2007; Dehghani et al. 2012; Khoo 2016 for some promising directions.)

5 Revision: A new puzzle for probabilistic counterfactuals and its solutions

The revision condition (13b) is meant to algorithmize the revised heuristic from sections 1-2: evaluate counterfactuals by holding fixed causally independent facts while revising facts that are causally downstream. By comparison to issues around backtracking, this aspect of counterfactual interpretation has been much less emphasized in theories of counterfactuals built around causal models. However, it is just as crucial: as Goodman (1947) already made clear, the key problematic in a theory of counterfactuals is to work out which aspects of the real world to hold fixed and which to vary in constructing a counterfactual scenario. This aspect of the interpretation of counterfactuals has, however, been a primary focus of inquiry in work in both similarity-based theories (Lewis 1979) and premise semantics (Kratzer 1981a, 1989; Veltman 1985, 2005, etc.). The revision condition (13b) is intended to play this role—and it seems possible to do without the additional formal machinery of similarity orderings or premise sets, at least when treating probabilistic

[8] The same held of matched indicative/counterfactual pairs with *likely*, *might*, *have to*, and *certain*, varying systematically information about the contents of the jars. For (17), the average slider rating was 84 out of 100, as compared to 81/100 for the matched indicative.

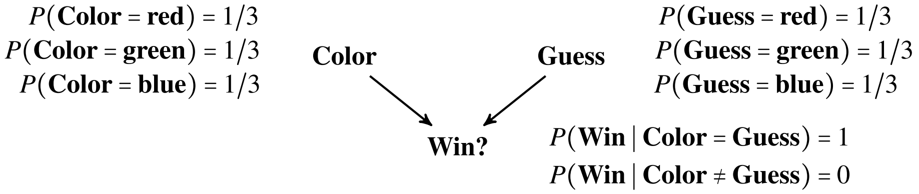


Figure 4 CBN for the guessing game.

counterfactuals. This would be a useful reduction, since there is abundant independent evidence for the cognitive importance of probabilistic reasoning and causal models. For discussion, see for example Glymour 2001; Sloman 2005; Gopnik & Schultz 2007; Griffiths, Kemp & Tenenbaum 2008; Tenenbaum, Kemp, Griffiths & Goodman 2011; Danks 2014 among many others.

But the simple revision algorithm in (13b) is, alas, too simple to capture the interplay between observational evidence, probabilities, and counterfactuals. Consider a guessing game: someone picks a ball at random from a jar containing equal numbers of red, blue, and green balls. You try to guess the color. If you get it right, you win. A CBN for this game is in Figure 4. Suppose that you guessed "blue". You are then told that you did not win, without learning which color did come up.

Against this background, the probabilistic counterfactual in (18) seems to be true.

(18) If you'd guessed "red", there's an exactly 50% chance that you'd have won.

The trouble is that the revision condition (13b) makes this counterfactual false. $\mathcal{O}$ is $\{\textbf{Guess} = \textbf{blue}, \textbf{Color} \neq \textbf{blue}\}$, and the antecedent intervenes on **Guess?**, leading to the removal of observations about both **Guess?** and **Color?** in the counterfactual scenario. $\mathcal{O}*$ then contains just the antecedent information **Guess = red**, and (18) is evaluated as

$$[\![\text{If you'd guessed "red", there is an exactly 50\% chance you'd have won}]\!]^{w,B,\mathcal{O}=\{\textbf{Guess=blue},\textbf{Color}\neq\textbf{blue}\}}$$

$$\begin{aligned}
&= [\![\text{there is an exactly 50\% chance you'd have won}]\!]^{w,B,\mathcal{O}=\{\textbf{Guess=red}\}}\\
&= \text{True iff } P_B(\textbf{Win} \mid \textbf{Guess} = \textbf{red}) = 1/2\\
&= \text{True iff } \sum_{C\in\{\textbf{red},\textbf{green},\textbf{blue}\}} P_B(\textbf{Win} \mid \textbf{Guess} = \textbf{red}, \textbf{Color} = C) \times P(\textbf{color} = C) = 1/2\\
&= \text{True iff } 1 \times 1/3 + 0 \times 1/3 + 0 \times 1/3 = 1/2\\
&= \text{True iff } 1/3 = 1/2\\
&= \text{False}
\end{aligned}$$

When evaluating the consequent of (18), (13b) tells us to ignore the information that I did not win, since it is causally downstream of the antecedent (which triggers an intervention on **guess**). Intuitively, though, the information that we did not actually win is needed to get the right result, because it is what allows us to infer that the ball was not blue. We seem to require a way to propagate the information that the ball is not blue throughout the CBN *before* discarding this information when constructing the counterfactual model. If we could do this, then the probability of **Color = red** and **Color = green** would both shift from 1/3 to 1/2. An improved version of (13b) should allow us to incorporate into our model diagnostic information about the actual values of causally independent variables via conditioning, before we proceed to ignore their actual values when evaluating the effects of an intervention.

What we would like is for the probability distribution on variables causally independent of the antecedent to be informed by *all* relevant observations. In particular, the fact that you did not win should be considered, along with your actual guess, in deciding what the probability is that the color is red. So there should be some opportunity for information to flow from the real-world observation $\overline{\textbf{win}}$ to flow to **Color**. However, we obviously do not want the observation $\overline{\textbf{win}}$ to persist in the counterfactual scenario that we construct when evaluating (18). If it were, the probability of **win** would be 0, not .5, and the sentence would be trivially false.

5.1 First solution: Recursive specification of counterfactual probabilities

A solution that requires only minor modification to our earlier definitions involves defining a counterfactual probability distribution piecemeal—call it P^*. In P^*, variables causally independent of the antecedent have the values we would expect from conditioning on the full observation set $\mathcal{O}$. Causally dependent variables are given a more complicated definition that ignores certain real-world information, as determined by the structure of the causal graph, while taking into account the information added to the causally independent part in the shift from P to P^*.

A compositional implementation of this idea is available, though slightly more complex than the semantics presented earlier. We have to represent the probability measure explicitly as a shiftable parameter P which is manipulated this measure on the observation set. (Here 'ϕ' abbreviates $[\![\phi]\!]^{w,B,\mathcal{O},P}$.)

(19) $[\![\text{If } \phi, \text{then } Op\ \psi]\!]^{w,B,\mathcal{O},P} = [\![Op\ \psi]\!]^{w,B,\mathcal{O}^+,P^+}$, where $\mathcal{O}^+ = \mathcal{O} \cup \{\phi\}$ and $P^+ = P(\cdot \mid \mathcal{O}^+)$.

(20) $[\![\text{If were } \phi, \text{would } Op\ \psi]\!]^{w,B,\mathcal{O},P} = [\![Op\ \psi]\!]^{w,B^*,\mathcal{O}^*,P^*}$ (where B^* and $\mathcal{O}^*$ are defined as above).

This definition is given with the possibility of embedded conditionals in mind. In general, we must also require that the parameter P is initialized to $P_B(\cdot \mid \mathcal{O})$ in unembedded sentences, in order to establish the expected relationship between the causal model B, the observation set $\mathcal{O}$, and the probability measure P.

The definitions of probability operators must also be modified to make reference to the probability measure parameter. For example:

(21) $[\![\text{likely}]\!]^{w,B,\mathcal{O},P} = \lambda d_d \lambda q_{\langle s,t \rangle} \,.\, P(q) > d$

By the stipulation just mentioned, unembedded *likely* will be interpreted relative to $P_B(\cdot \mid \mathcal{O})$ as in our earlier definition. However, the modified definition allows us probability operators to track the effects of compositionally modifying the probability measure and the observation set independently.

The remaining question is how to define P^* for counterfactuals. One way to implement the idea sketched informally above is to begin by defining P^* from the parameter P only for variables causally independent of the antecedent.

$$P^*(V) = P(V \mid \mathcal{O}) \text{ if } V \text{ is causally independent of the antecedent}$$

The causally independent part of the counterfactual distribution takes into account all evidence available (though not, crucially, the antecedent). For example, $P^*(\textbf{Color} = \textbf{red})$ is correctly predicted to be $1/2$, since it is equal to $P_B(\textbf{Color} = \textbf{red} \mid \mathcal{O})$—taking into account the observation **win**. Since $\mathcal{O}$ also includes $\textbf{Guess} = \textbf{blue}$ and the model specifies $P(\textbf{Win} \mid \textbf{Color} = \textbf{Guess}) = 1$, we can infer $P(\textbf{Color} \neq \textbf{Guess} \mid \mathcal{O}) = 1$, and so $P(\textbf{red} \mid \mathcal{O}) = P(\textbf{green} \mid \mathcal{O})$, with both normalized to $1/2$. Since **Color** is not downstream of the antecedent **Guess**, we have $P^*(\textbf{red}) = P_B(\textbf{red} \mid \mathcal{O}) = 1/2$, as per intuition.

Next we extend P^* to variables causally dependent of the antecedent, including the antecedent variable itself. The intuition behind the definition is that the counterfactual distribution for these variables should not be conditioned directly on $\mathcal{O}$, but it should take into account the update to other variables induced by the first part of the definition, together with the antecedent and the stable information in the conditional probability tables. (Recall that $\mathcal{O}^*$ is $\mathcal{O}$ minus any value for any variable causally dependent on the antecedent, plus the antecedent.)

$$P^*(V) = P(V \mid \mathit{Parents}(V), \mathcal{O}^*) \times P^*(\mathit{Parents}(V)) \text{ if } V \text{ is downstream of the antecedent}$$

This definition propagates information recursively down from the causally independent part to the dependent part via (i) the updated distribution $P^*(\mathit{Parents}(V))$ and (ii) the constraints on $P(V \mid \mathit{Parents}(V))$ given by the conditional probability tables, while also taking into account (iii) the constraint imposed by the antecedent (which is contained in $\mathcal{O}^*$).

Given this, we evaluate the key example as follows. *If you'd guessed "red", there is an exactly 50% chance you'd have won* is true iff $P^*(\textbf{Win?} = \text{True}) = .5$ in the counterfactual scenario constructed on the basis of the antecedent *If you'd guessed "red"*. To check whether this equality holds, we first compute the values of P^* for all causally independent variables in the model in Figure 4—here, just **Color**. We did this above, finding $P^*(\textbf{red}) = .5$. Next we propagate counterfactual probabilities to the causally dependent variables. (Since all observations in $\mathcal{O}$ are causally dependent on the antecedent, $\mathcal{O}^*$ contains only the antecedent $\textbf{Guess} = \textbf{red}$.)

$$\begin{aligned} P^*(\textbf{Guess} = \textbf{red}) &= P(\textbf{Guess} = \textbf{red} \mid \mathit{Parents}(\textbf{Guess}), \mathcal{O}^*) \times P^*(\mathit{Parents}(\textbf{Color})) \\ &= P(\textbf{Guess} = \textbf{red} \mid \textbf{Guess} = \textbf{red}) \qquad\qquad [\text{since } \textbf{Guess} \text{ has no parents}] \\ &= 1 \end{aligned}$$

With this value in hand for $P^*(\textbf{Guess} = \textbf{red})$ we can find the target value, the counterfactual probability of $\textbf{Win?} = \mathit{True}$.

$$P^*(\textbf{Win?} = \text{True}) = P(\textbf{Win?} = \text{True} \mid \mathit{Parents}(\textbf{Win?}), \mathcal{O}^*) \times P^*(\mathit{Parents}(\textbf{Win?}))$$

Unpacking the references to $\mathit{Parents}(\textbf{Win})$ and recalling that $\mathcal{O}^* = \{\textbf{Guess} = \textbf{red}\}$, this is equivalent to

$$P^*(\textbf{Win?} = \text{True}) = \sum_{C \in \{\textbf{red, blue, green}\}} P(\textbf{Win?} = \text{True} \mid \textbf{Guess} = \textbf{red}, \textbf{Color} = C) P^*(\textbf{Color} = C)$$

where *Colors* is $\{\textbf{red, blue, green}\}$.

This is a sum over the following three components:

- **red**: $P(\textbf{Win?} = \text{True} \mid \textbf{Guess} = \textbf{red}, \textbf{Color} = \textbf{red}) \times P^*(\textbf{Color} = \textbf{red}) = 1 \times .5 = .5$
- **blue**: $P(\textbf{Win?} = \text{True} \mid \textbf{Guess} = \textbf{blue}, \textbf{Color} = \textbf{red}) \times P^*(\textbf{Color} = \textbf{blue}) = 0 \times 0 = 0$
- **green**: $P(\textbf{Win?} = \text{True} \mid \textbf{Guess} = \textbf{green}, \textbf{Color} = \textbf{red}) \times P^*(\textbf{Color} = \textbf{green}) = 0 \times .5 = 0$

So, the result is $P^*(\textbf{Win?} = \text{True}) = .5$, which was the intuitively correct answer.

This method of deriving counterfactual probabilities was chosen for two reasons. First, it gets the right result for the guessing game example. Second, it does this while maintaining as closely as possible the idea that we argued for earlier: that, when we are constructing counterfactual scenarios, real-world information about variables causally downstream of the antecedent is not taken into account in the same way as information about causally independent variables.

Of course, it is necessary to check that the modified proposal does not disrupt the accounts of other puzzles discussed earlier. For our solution to the probabilistic Morgenbesser example (16), the revised definition affects the reasoning process slightly but not the outcome—the counterfactual remains false. (Recall that $\mathcal{O} = \{\textbf{Heads}, \overline{\textbf{Accept Bet}}, \overline{\textbf{Win}}\}$ and $\mathcal{O}^* = \{\textbf{Heads}, \textbf{Accept Bet}\}$ in this example.)

$$\begin{aligned}
&[\![\text{If you had bet, there is an exactly 50\% chance that you would have won}]\!]^{w,B,\mathcal{O},P_B(\cdot|\mathcal{O})} \\
&\qquad\qquad = [\![\text{there is an exactly 50\% chance that you would have won}]\!]^{w,B^*,\mathcal{O}^*,P^*} \\
&\qquad\qquad = \text{True iff } P^*(\textbf{Win}) = .5
\end{aligned}$$

The counterfactual distribution on the causally independent variable **Heads?** is given by $P^*(\textbf{Heads}) = P(\textbf{Heads} \mid \mathcal{O})$. Since **Heads** is in $\mathcal{O}$, P^* assigns probability 1 to heads and 0 to tails. (Indeed, for any causally independent variable whose value is observed, the revised interpretation of counterfactuals using P^* will agree with the probability assigned by the earlier, simpler definition.)

For the variables **Accept bet?** and **Win?**—which are causally dependent on the antecedent—$P\star$ is derived by conditioning on the variable's parents, the pruned observations, and the antecedent. $P^*(\textbf{Accept bet})$ is of course 1, and so

$$\begin{aligned}
P^*(\textbf{Win?} = \text{True}) &= \sum_{v \in \{\textbf{Heads, Tails}\}} \; \sum_{v' \in \{\textbf{Accept, Reject}\}} P(\textbf{win} \mid v, v') \times P^*(v, v') \\
&= \sum_{v \in \{\textbf{Heads, Tails}\}} \; \sum_{v' \in \{\textbf{Accept, Reject}\}} P(\textbf{win} \mid v, v') \times P^*(v) \times P^*(v')
\end{aligned}$$

(where the second line follows because **Heads?** and **Accept bet?** are probabilistically independent). Since $P^*(\textbf{Accept}) = 1$ and $P^*(\textbf{Heads}) = 1$, the only non-zero value in this summation is

$$P(\textbf{Win?} = \text{True} \mid \textbf{Heads}, \textbf{Accept}) \times P^*(\textbf{Heads}) \times P^*(\textbf{Accept}),$$

$P(\textbf{Win?} = \text{True} \mid \textbf{Heads}, \textbf{Accept})$ is 1 by the conditional probability table in Figure 2, corresponding to the scenario's set-up: you win only if you accept the bet, and the coin comes up heads. So, $P^*(\textbf{Win})$ is 1 as well, and the probabilistic Morgenbesser counterfactual is correctly ruled false in the revised definition.

While I will not go through the derivations in detail, it is straightforward to check that the revised definitions do not affect our earlier results for the flight examples. As in the probabilistic Morgenbesser example, the agreement between the earlier and revised definitions is due to the fact that all variables in the model are observed. In contrast, the crucial property of the guessing game example that led us to complicate our definitions was that one relevant variable—**Color**—was not observed.

5.2 Second solution: Structural Causal Models

While the revised definition of counterfactual probabilities for CBNs gets the right result for the guessing game puzzle, one might find the piecemeal style of definition less than fully elegant. In addition, we have not explored the predictions of the new definition beyond a handful of simple examples, and it remains to be seen whether further cases will motivate a more complex or even totally different approach. In any case, it is worth developing an alternative solution that also gets the right answer, and may turn out to yield different predictions in more complex examples.

The alternative solution relies on a different, but closely related, representation of causation and uncertainty—the Structural Causal Models (SCMs) associated most prominently with the work of Pearl (2000, 2009). (See also Kline 2016: §8; Pearl & Mackenzie 2018 for introductory presentations.) At a high level, the key difference between CBNs and SCMs is that the latter factorize the information in a CBN's conditional probability tables into two components. The first is a set of exogenous (causeless) sources of randomness E_V, one for each ordinary variable V. We can think of an exogenous random variable either as a random number generator or as a summary of additional, independent causal factors that are not included in the model. The second component is a deterministic function f that takes as input E_V as well the values of V's parents in the corresponding CBN. Figure 5 shows what the guessing game would look like under this factorization. The model captures the same real-world information about the game as the CBN in Figure 4, but does so in a more complicated way whose usefulness in counterfactual interpretation will become apparent shortly.

Formally, a SCM can be defined as follows. Comparison to the definition of a CBN in (7) makes it clear that the key difference is in the factorization of the information in a CBN's conditional probability tables into a random component $\mathcal{E}$ and a deterministic component $\mathcal{F}$.

(22) A structural causal model C is given by $\langle W, \mathcal{Q}, \mathcal{A}, \mathcal{E}, P, \mathcal{F} \rangle$, where

a. W is a set of possible worlds,

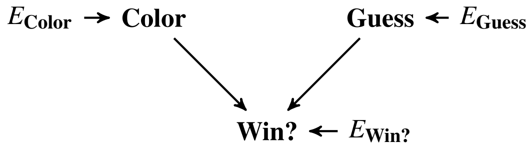
b. $\mathcal{Q}$ is a set of regular ("endogenous") variables,

c. $\mathcal{A}$ is an acyclic binary relation on $\mathcal{Q}$ (the "arrows"),

d. $\mathcal{E} = \{E_Q\}_{Q \in \mathcal{Q}}$ is a set of parentless ("exogenous") random variables,

e. P is a prior probability distribution over $\mathcal{E}$ in which all variables in $\mathcal{E}$ are independent.[9]

[9] P can be extended to the endogenous $Q \in \mathcal{Q}$ by treating the probability of $Q = q$ as the total probability of all ways of providing values to the exogenous $\mathcal{E}$ that force $q = Q$ via the functions $\mathcal{F}$ defined below.



$$\forall C \in \{\textbf{red}, \textbf{green}, \textbf{blue}\} : P(E_{\textbf{Color}} = C) = 1/3$$
$$\forall C \in \{\textbf{red}, \textbf{green}, \textbf{blue}\} : P(E_{\textbf{Guess}} = C) = 1/3$$
$$\textbf{Guess} = E_{\textbf{Guess}}$$
$$\textbf{Color} = E_{\textbf{Color}}$$
$$\textbf{Win?} = \text{True iff } (\textbf{Color} = \textbf{Guess})$$

Figure 5 Structural Causal Model for the guessing game.

f. $\mathcal{F} = \{f_Q\}_{Q \in \mathcal{Q}}$ is a set of functions assigning to each endogenous variable $Q \in \mathcal{Q}$ one of its cells $q \in Q$ as a function of (i) the value of the associated exogenous random variable E_Q, and (ii) a (possibly empty) vector $\vec{\mathbf{v}}$ of values for Q's parents in $\mathcal{A}$, if any.

The $f_Q \in \mathcal{F}$ operate like meaning postulates in Montague semantics: a possible world w—or, if you like, a vector of values for the variables in $\mathcal{Q}$—can be represented only if the value of Q at w (the cell of Q containing w) matches the output of f_Q, given as arguments (i) some possible value of the random variable E_Q, and (ii) the values of the Q's parent variables in w.

Note that the appearance of deterministic functions in this model does not imply that the model can only be applied in situations where causation is genuinely deterministic. For his part, Pearl (2000) insists on an interpretation where causation is genuinely deterministic at the macro-level, and argues that this feature accurately models humans' intuitive thinking about causation. On this interpretation the randomness of the variables in $\mathcal{E}$ is merely a reflection of our ignorance about the myriad real-world factors which may be responsible for determining their values.

However, it is equally consistent with the formal definition of a SCM to interpret the exogenous variables as encoding true randomness. For instance, if the value of $E_{\textbf{Color}}$ in Figure 4 is genuinely non-deterministic, then the value of **Color** will be non-deterministic with the same (extended) probability distribution. Indeed, the factorization of information in SCMs is a generalization of the familiar factorization of linear statistical models into a deterministic part and a stochastic error term, as in $y = 3x + \mathcal{N}(0,1)$: the randomness of the Gaussian error term trickles through the deterministic function to yield something that can, if desired, be interpreted as a genuinely non-deterministic mapping from x to y. The factorization of information in SCMs can also be seen as an instance of the way that random functions are defined in probabilistic programming languages, as deterministic functions one of whose arguments is the output of a random number generator: see, for example, Goodman, Mansinghka, Roy, Bonawitz & Tenenbaum 2008; Pfeffer 2016.

In addition to these two interpretations, the content of the exogenous variables in a SCM could also be interpreted as summarizing a mixture of truly random information and unmodeled real-world factors. In any case, the key point here is that scholars who are skeptical about deterministic causation need not reject the use of SCMs simply because of the appearance of deterministic functions in their definition.

Using SCMs, we can modify the definition of counterfactuals (13) in order to account for the troublesome guessing game example (18). We now relativize interpretation to a SCM C rather than a CBN B. Crucially, the interpretation of counterfactuals involves conditioning the exogenous variables on **all** real-world information, before selectively discarding values of endogenous variables that are incompatible with the counterfactual supposition given the constraints imposed by $\mathcal{F}$.[10]

(23) $[\![\text{If were } \phi\text{, would } Op\ \psi]\!]^{w,C,\mathcal{O}} = [\![Op\ \psi]\!]^{w,C^*,\mathcal{O}}$, where $C^* = \langle W, \mathcal{Q}, \mathcal{A}^*, \mathcal{E}, P^*, \mathcal{F}^* \rangle$ modifies $C = \langle W, \mathcal{Q}, \mathcal{A}, \mathcal{E}, P, \mathcal{F} \rangle$ so that

- $\mathcal{A}^*$ is $\mathcal{A}$ minus any arrows pointing into $\mathcal{Q}_\phi$, the question that ϕ addresses;
- P^* is $P(\cdot \mid \mathcal{O})$—conditioned on the full observation set (!!);
- $\mathcal{F}^*$ is $\mathcal{F}$ with the equation determining $\mathcal{Q}_\phi$'s value replaced with the constant function $\mathcal{Q}_\phi = [\![\phi]\!]^{w,C,\mathcal{O}}$.

To see how this works, consider the SCM of the guessing game in Figure 5. Here $C = \langle W, \mathcal{Q}, \mathcal{A}, \mathcal{E}, P, \mathcal{F} \rangle$, where

- W contains at least one world in each cell of the maximally fine-grained question $\sqcap \mathcal{Q}$,
- $\mathcal{Q} = \{\textbf{Color}, \textbf{Guess}, \textbf{Win?}\}$,
- $\mathcal{A} = \{\langle \textbf{Color}, \textbf{Win?} \rangle, \langle \textbf{Guess}, \textbf{Win?} \rangle\}$,
- $\mathcal{E} = \{E_{\textbf{Color}}, E_{\textbf{Guess}}, E_{\textbf{Win?}}\}$,
- P satisfies the constraints on the right of Figure 5 while also rendering all elements of $\mathcal{E}$ independent[11],
- $\mathcal{F} = \{f_{\textbf{Color}}, f_{\textbf{Guess}}, f_{\textbf{Win?}}\}$, where
  i. $f_{\textbf{Color}}(E_{\textbf{Color}}) = E_{\textbf{Color}}$,
  ii. $f_{\textbf{Guess}}(E_{\textbf{Guess}}) = E_{\textbf{Guess}}$,
  iii. $f_{\textbf{Win?}}(E_{\textbf{Win?}}, \textbf{Color}, \textbf{Guess}) = \text{True iff } \textbf{Color} = \textbf{Guess}$.

The troublesome example is repeated here.

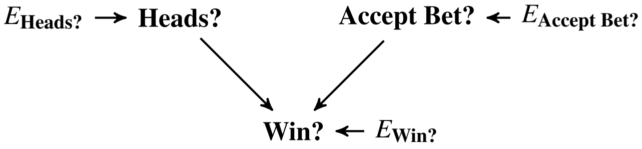
(24) [You guessed "blue", and you didn't win.] If you'd guessed "red", there's an exactly 50% chance that you'd have won.

Definition (23) instructs us to interpret (24) as equivalent to its consequent evaluated against a modified SCM $C^* = \langle W, \mathcal{Q}, \mathcal{A}^*, \mathcal{E}, P^*, \mathcal{F}^* \rangle$, with the starred components modified as follows.

- $\mathcal{A}^* = \mathcal{A}$ (since no arrows point into **Guess**),
- $P^* = P(\cdot \mid \mathcal{O}) = P(\cdot \mid \textbf{Guess} = \textbf{blue}, \textbf{Win?} = \text{False})$,
- $\mathcal{F}^* = \{f_{\textbf{Color}}, f_{\textbf{Guess}}, f_{\textbf{Win?}}\}$, where
  i*. $f^*_{\textbf{Color}}(E_{\textbf{Color}}) = E_{\textbf{Color}}$, (unchanged)

[10] The definition assumes that the consequent ψ does not make reference to any exogenous variables, which are generally assumed to be unavailable for explicit mention or observation.

[11] The definition of SCMs requires that $E_{\textbf{Win?}}$ exist, but its distribution has no effect on the behavior of the model since it is ignored by $f_{\textbf{Win?}}$. As a result I have not specified a distribution for $P(E_{\textbf{Win?}})$.



$$P(E_{\textbf{Heads?}} = \textbf{heads}) = .5$$
$$P(E_{\textbf{Heads?}} = \textbf{tails}) = .5$$
$$\textbf{Heads} = E_{\textbf{Heads?}}$$
$$\textbf{Accept Bet?} = E_{\textbf{Accept Bet?}}$$
$$\textbf{Win?} = \text{True iff } (\textbf{heads} \cap \textbf{accept})$$

Figure 6 SCM for the probabilistic Morgenbesser example (4).

ii*. $f^*_{\textbf{Guess}}(E_{\textbf{Guess}}) = \textbf{red}$, (constant function yielding **red**)

iii*. $f^*_{\textbf{Win?}}(E_{\textbf{Win?}}, \textbf{Color}, \textbf{Guess}) = \text{True iff } \textbf{Color} = \textbf{Guess}$. (unchanged)

In the modified model $f_{\textbf{Guess}}$ ignores its arguments, and is set to a constant value corresponding to the antecedent. This is the locus of the intervention [Pearl's "action" step, associated with his *do* operator]. The other key effect of counterfactual interpretation is to condition P on the full observation set $\mathcal{O}$ to yield $P(\cdot \mid \textbf{Guess} = \textbf{blue}, \textbf{Win?} = \text{False})$ [Pearl's somewhat ill-named "abduction" step]. Given the observation **Guess = blue** and the original model's deterministic requirement $\textbf{Guess} = E_{\textbf{Guess}}$, the updated distribution will yield $P^*(E_{\textbf{Guess}} = \textbf{blue}) = 1$. In addition, given $\textbf{Win?} = \text{False}$ we infer $\textbf{Color} \neq \textbf{blue}$, and so $E_{\textbf{Color}} \neq \textbf{blue}$. Conditioning on the observed values of **Guess** and **Win?** thus yields

$$\begin{aligned} P(E_{\textbf{Color}} = \textbf{red} \mid \textbf{Guess} = \textbf{blue}, \textbf{Win?} = \text{False}) &= P(E_{\textbf{Color}} = \textbf{red} \mid E_{\textbf{Color}} \neq \textbf{blue}) \\ &= \frac{P(E_{\textbf{Color}} = \textbf{red})}{P(E_{\textbf{Color}} = \textbf{red}) + P(E_{\textbf{Color}} = \textbf{green})} \\ &= \frac{1/3}{1/3 + 1/3} \\ &= .5 \end{aligned}$$

Given this updated distribution, we can now apply the modified causal constraints $\mathcal{F}$ to propagate probabilities throughout the deterministic part of the model. Since $P(E_{\textbf{Color}} = \textbf{red}) = .5$ and function (i*) requires $\textbf{Color} = E_{\textbf{Color}}$, we have $P(\textbf{Color} = \textbf{red}) = .5$. Since $\textbf{Guess} = \textbf{red}$ as a result of our intervention (modified constraint (ii*)) and we win iff $\textbf{Color} = \textbf{Guess}$, **Win** is also true with probability .5. As a result, (18) comes out true, as desired, according to the revised definition (23) that made crucial use of the factorization of information in SCMs.

Again, before declaring victory we must check that the shift to SCMs does not disrupt our earlier accounts of the probabilistic Morgenbesser example and the flight examples. For the Morgenbesser example *If you'd bet there is an exactly .5 chance you'd have won*, Figure 6 specifies a SCM corresponding to the CBN in Figure 3. Following the interpretive procedure described above, we update P to $P(\cdot \mid \textbf{Heads}, \overline{\textbf{accept}}, \overline{\textbf{win}})$, and then intervene by replacing the function determining **Accept Bet?** with the constant function **Accept Bet?** = accept. Since **Win?** = True if and only if **heads** and **accept** are both true, and **accept** is always true in the counterfactual model, the probability of winning is equal to the probability of heads. The latter value is just

$$P(\mathbf{heads} \,|\, \mathbf{heads}, \overline{\mathbf{accept}}, \overline{\mathbf{win}}),$$

which is obviously 1. So, the SCM-based definition returns the same result as the CBN-based definition also for the probabilistic Morgenbesser example (16): the sentence is false because the true counterfactual probability is 1. The reader may wish to verify that a similar treatment yields the same result as our earlier interpretive schemata when the flight examples are translated into the format of SCMs.

6 Conclusion

This paper has analyzed the interaction of probabilistic and causal information in the interpretation of both indicative and counterfactual conditionals. The key result is that the combination of causal models and a restrictor syntax makes it possible to give probabilistic conditionals of both types a unified semantic treatment. The difference between indicatives and counterfactuals, on this account, reduces to the choice between conditioning—which adds information monotonically—and intervention—which modifies the causal model and removes certain information.

The first part of the paper described and motivated a simple algorithm for constructing counterfactual probability distributions: simply ignore information about observations that are causally dependent on the antecedent. This approach makes correct predictions in several cases and does much to motivate the use of causal models in interpreting counterfactuals—as opposed, for example, to more general methods of counterfactual revision that do not explain why causal relevance and directionality have the specific effects that they do on the interpretation of counterfactuals (e.g., Kratzer 1981b; Veltman 2005). However, the first-pass algorithm fails in more complex scenarios in which some variables that are causally independent of the antecedent are not observed. For these variables, we need to update probabilities on the basis of all available observations—including those of variables dependent on the antecedent—before using this information to create a counterfactual distribution that includes variables causally dependent on the antecedent. I discussed two ways of accomplishing this: a novel algorithm for constructing counterfactual probability distributions in causal Bayes nets, and the existing proposal of Pearl (2000, 2009) that relies on structural causal models. Both of these approaches get the right result for the puzzle at hand, but they may well make divergent predictions for more complex cases. An important outstanding question is whether and in what cases the two approaches to interpreting counterfactual probabilities truly differ.

This approach to the interpretation of conditionals also has important limitations. For one, I have not specified a procedure for interpreting non-probabilistic conditionals. Bare conditionals are a general puzzle for restrictor analyses of conditionals, where the antecedent's semantic effect is restricted to its effect on the interpretation of operators in the consequent. Some available options include the assumption of a silent epistemic operator in the consequent (as in Kratzer 1991a,b) or an account where bare conditionals receive probability values (cf. (Stalnaker & Jeffrey 1994; Kaufmann 2004, 2009)). While it is not necessary to choose among these options here, the prospects of the account proposed here do depend on the eventual viability of a proposal along these lines.

In addition, we have limited attention to sentences in which the antecedent of the conditional was a complete answer to a unique question Q (value of a unique variable) in the causal model. In the more general case—for example, when the antecedent is logically complex—the interpretation will have to be more complicated. This is a very general problem for theories of counterfactuals built around causal models, which are generally tailored for the case where each intervention is the assignment of a value to a variable, or a conjunction of such assignments. While there is still much work to be done here, several proposals exist that can be combined directly with the semantics given here: see Briggs 2012; Ciardelli et al. 2018; Lassiter 2017a.

References

Adams, Ernest W. 1975. *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Springer.

Arregui, Ana Cristina. 2005. *On the accessibility of possible worlds: The role of tense and aspect*: University of Massachusetts PhD dissertation.

Barker, Stephen. 1998. Predetermination and tense probabilism. *Analysis* 58(4). 290–296.

Barker, Stephen. 1999. Counterfactuals, probabilistic counterfactuals and causation. *Mind* 108(431). 427–469.

Belnap, Nuel & Mitchell Green. 1994. Indeterminism and the thin red line. *Philosophical perspectives* 8. 365–388.

Briggs, Rachael. 2012. Interventionist counterfactuals. *Philosophical studies* 160(1). 139–166.

Ciardelli, Ivano, Linmin Zhang & Lucas Champollion. 2018. Two switches in the theory of counterfactuals: A study of truth conditionality and minimal change. *Linguistics and Philosophy* .

Danks, David. 2014. *Unifying the Mind: Cognitive Representations as Graphical Models*. MIT Press.

Dehghani, Morteza, Rumen Iliev & Stefan Kaufmann. 2012. Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language* 27(1). 55–85.

Edgington, Dorothy. 1995. On conditionals. *Mind* 104(414). 235–329.

Edgington, Dorothy. 2003. Counterfactuals and the benefit of hindsight. In Phil Dowe & Paul Noordhof (eds.), *Cause and Chance: Causation in an Indeterministic World*, Routledge.

Edgington, Dorothy. 2008. Counterfactuals. In *Proceedings of the Aristotelian Society* 108 1, 1–21.

Edgington, Dorothy. 2011. Causation first: Why causation is prior to counterfactuals. In Christoph Hoerl, Teresa McCormack & Sarah R. Beck (eds.), *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology*, 230–241. Oxford University Press.

Fine, Kit. 1975. Vagueness, truth and logic. *Synthese* 30(3). 265–300.

Gerstenberg, Tobias, Christos Bechlivanidis & David A Lagnado. 2013. Back on track: Backtracking in counterfactual reasoning, .

Glymour, Clark N. 2001. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT press.

Goodman, N.D., V.K. Mansinghka, D. Roy, K. Bonawitz & J.B. Tenenbaum. 2008. Church: a language for generative models. In *Uncertainty in Artificial Intelligence* 22, 23.

Goodman, Nelson. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy* 44(5). 113–128.

Gopnik, Alison & Laura Schultz (eds.). 2007. *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press.

Griffiths, Thomas L., Charles Kemp & Joshua B. Tenenbaum. 2008. Bayesian models of cognition. In Ron Sun (ed.), *Cambridge Handbook of Computational Psychology*, 59–100. Cambridge University Press.

Hiddleston, Eric. 2005. A causal theory of counterfactuals. *Noûs* 39(4). 632–657.

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

Kaufmann, Stefan. 2001a. *Aspects of the meaning and use of conditionals*: Stanford PhD dissertation.

Kaufmann, Stefan. 2001b. Tense probabilism properly conceived. In *Thirteenth Amsterdam Colloquium*, 132–137.

Kaufmann, Stefan. 2004. Conditioning against the grain. *Journal of Philosophical Logic* 33(6). 583–606.

Kaufmann, Stefan. 2005. Conditional predictions: A probabilistic account. *Linguistics and Philosophy* 28(2). 181–231. doi:10.1007/s10988-005-3731-9.

Kaufmann, Stefan. 2009. Conditionals right and left: Probabilities for the whole family. *Journal of Philosophical Logic* 38(1). 1–53. doi:10.1007/s10992-008-9088-0.

Kaufmann, Stefan. 2013. Causal premise semantics. *Cognitive science* 37(6). 1136–1170.

Khoo, Justin. 2015. On indicative and subjunctive conditionals. *Philosopher's Imprint* 15(32).

Khoo, Justin. 2016. Backtracking counterfactuals revisited. *Mind* .

Kline, Rex B. 2016. *Principles and practice of structural equation modeling*. Guilford publications 4th edn.

Kratzer, Angelika. 1981a. The notional category of modality. In Eikmeyer & Rieser (eds.), *Words, Worlds, and Contexts*, 38–74. de Gruyter.

Kratzer, Angelika. 1981b. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10(2). 201–216.

Kratzer, Angelika. 1989. An investigation of the lumps of thought. *Linguistics and philosophy* 12(5). 607–653.

Kratzer, Angelika. 1991a. Conditionals. In A. von Stechow & D. Wunderlich (eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*, 651–656. Walter de Gruyter.

Kratzer, Angelika. 1991b. Modality. In Arnim von Stechow & Dieter Wunderlich (eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*, 639–650. Walter de Gruyter.

Kvart, Igal. 1986. *A theory of counterfactuals*. Hackett.

Kvart, Igal. 1992. Counterfactuals. *Erkenntnis* 36(2). 139–179.

Lassiter, Daniel. 2010. Gradable epistemic modals, probability, and scale structure. In Nan Li & David Lutz (eds.), *Semantics & Linguistic Theory (SALT) 20*, 197–215. CLC Publications.

Lassiter, Daniel. 2015. Epistemic comparison, models of uncertainty, and the disjunction puzzle. *Journal of Semantics* 32(4). 649–684.

Lassiter, Daniel. 2017a. Complex antecedents and probabilities in causal counterfactuals. In Alexandre Cremers, Thom van Gessel & Floris Roelofsen (eds.), *21st Amsterdam Colloquium*, 45–54.

Lassiter, Daniel. 2017b. *Graded Modality*. Oxford University Press.

Lassiter, Daniel. 2018. Talking about (quasi-)higher-order uncertainty. In Cleo Condoravdi & Tracy Holloway King (eds.), *Tokens of Meaning: Papers in Honor of Lauri Karttunen*, CSLI Publications.

Lewis, David. 1973. *Counterfactuals*. Harvard University Press.

Lewis, David. 1979. Attitudes *de dicto* and *de se*. *The philosophical review* 88(4). 513–543.

Lucas, Christopher G. & Charles Kemp. 2015. An improved probabilistic account of counterfactual reasoning. *Psychological Review* 122(4). 700–734.

MacFarlane, John. 2003. Future contingents and relative truth. *The philosophical quarterly* 53(212). 321–336.

Meek, Christopher & Clark Glymour. 1994. Conditioning and intervening. *The British journal for the philosophy of science* 45. 1001–1021.

Moss, Sarah. 2015. On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics* 8. 1–81.

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

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press 2nd edn.

Pearl, Judea & Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pfeffer, Avi. 2016. *Practical Probabilistic Programming*. Manning Publications Co.

Rips, L. 2010. Two causal theories of counterfactual conditionals. *Cognitive science* 34(2). 175–221.

Santorio, Paolo. 2016. Interventions in premise semantics. *Philosophers' Imprint* .

Schaffer, Jonathan. 2004. Counterfactuals, causal independence and conceptual circularity. *Analysis* 64(284). 299–308.

Schulz, Katrin. 2007. *Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals*: ILLC, University of Amsterdam PhD dissertation.

Schulz, Katrin. 2011. "If you'd wiggled A, then B would've changed": Causality and counterfactual conditionals. *Synthese* 179(2). 239–251.

Sloman, Steven A. 2005. *Causal Models: How We Think About the World and its Alternatives*. OUP.

Sloman, Steven A & David A Lagnado. 2005. Do we "do"? *Cognitive Science* 29(1). 5–39.

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

Stalnaker, Robert. 2015. Counterfactuals and humean reduction. *A Companion to David Lewis* 57. 411.

Stalnaker, Robert & Richard Jeffrey. 1994. Conditionals as random variables. In *Probability and conditionals: Belief revision and rational decision*, 31–46. Cambridge University Press.

Swanson, Eric. 2006. Interactions With Context. Ph.D. thesis, MIT.

Swanson, Eric. 2011. How not to theorize about the language of subjective uncertainty.

Swanson, Eric. 2015. The application of constraint semantics to the language of subjective uncertainty. *Journal of Philosophical Logic* 1–26.

Tenenbaum, Joshua B., Charles Kemp, Tom L. Griffiths & Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022). 1279–1285.

Tichý, Pavel. 1976. A counterexample to the stalnaker-lewis analysis of counterfactuals. *Philosophical Studies* 29(4). 271–273.

Ülkümen, Gülden, Craig R Fox & Bertram F Malle. 2015. Two dimensions of subjective uncertainty: Clues from natural language. To appear in *Journal of Experimental Psychology: General*.

Veltman, Frank. 1985. *Logics for conditionals*: University of Amsterdam PhD dissertation.

Veltman, Frank. 2005. Making counterfactual assumptions. *Journal of Semantics* 22(2). 159–180.

Yalcin, Seth. 2007. Epistemic modals. *Mind* 116(464). 983–1026.

Yalcin, Seth. 2010. Probability operators. *Philosophy Compass* 5(11). 916–937.

Yalcin, Seth. 2012. Context probabilism. In M. Aloni, V. Kimmelman, F. Roelofsen, G. W. Sassoon, K. Schulz & M. Westera (eds.), *Logic, Language and Meaning* Lecture Notes in Computer Science 7218, 12–21. Springer.

Zhao, Michael. 2015. Intervention and the probabilities of indicative conditionals. *The Journal of Philosophy* 112(9). 477–503.