Complex antecedents and probabilities in causal counterfactuals*

Daniel Lassiter[1]

Stanford University, Stanford, California, USA

Abstract

Ciardelli, Zhang, & Champollion [5] point out an empirical problem for theories of counterfactuals based on maximal similarity or minimal revision involving negated conjunctions in the antecedent. They also show that disjunctions and negated conjunctions behave differently in counterfactual antecedents, and propose an attractive solution that combines Inquisitive Semantics [4] with a theory of counterfactuals based on interventions on causal models [20]. This paper describes several incorrect empirical predictions of the resulting account, which point to a very general issue for interventionist theories: frequently the antecedent does not give us enough information to choose a unique intervention. The problem applies also to indefinites and to the negation of any non-binary variable. I argue that, when there are multiple ways of instantiating a counterfactual antecedent, we prefer scenarios that are more likely given general probabilistic causal knowledge. A theory is proposed which implements this idea while preserving [5]'s key contributions.

> If I were not a physicist, I would probably be a musician. I often think in music. I live my daydreams in music. I see my life in terms of music.
> — Albert Einstein

1 Introduction

Interventionist theories of counterfactuals have been prominent in recent years in computer science, philosophy of science, statistics, psychology, and many other fields. While many have contributed to this enterprise, Pearl's *Causality* [20] is the most influential document by far. Pearl proposes that counterfactual reasoning proceeds by mutating a model of the causal structure of the world to render the antecedent true, and then considering what follows by causal laws. Semanticists and philosophers of language have begun to explore this approach as well (e.g., [5, 9, 10, 11, 21, 22]). While very attractive, the interventionist semantics is not as well-developed for linguistic purposes as theories based on similarity [15] or premise sets [12, 24].

Most critical, perhaps, is the need to deal seriously with the problem of complex antecedents. If Einstein had said *If I were a musician ...*, the necessary intervention would be fairly clear: we mutate the causal model to make Einstein a musician, and observe what the effects of this change are. But the interventionist semantics does not tell us what to do with his daydream *If I were not a physicist ...*. The problem is just that there are too many alternative professions. When we mutate the causal model so that Einstein is not a physicist, should we make him a barber? an electrician? a musician? unemployed? How can we choose among this bewildering variety of options? Worse, what are we to make of the *probably* in the consequent—if we want Einstein's claim to come out true, do we somehow intervene *non-determistically*, making him

*Many thanks to Lucas Champollion and Thomas Icard for numerous conversations which helped me to get clearer on these issues. Thanks also to Ivano Ciardelli and audiences at UC Davis Language Sciences and the New York Philosophy of Language Workshop.

a musician *most* of the time but sometimes something else? Pearl's semantics is silent on these questions. Failure to treat complex antecedents imposes severe limits on the linguistic generality of the inteventionist approach. These restrictions may well be unproblematic for some modeling purposes, but they are not acceptable if the interventionist semantics is to be linguistically respectable—and to vie with accounts based on similarity or premise sets.

In a recent paper Ciardelli, Zhang, & Champollion [5]—henceforth "CZC"—make a number of important contributions to this problem. First, they show experimentally that negated conjunctions in the antecedent do not behave as expected under maximal similarity/minimal revision theories (see also [2]). Second, they demonstrate the value of the interventionist semantics by providing a natural extension of Pearl's semantics to complex Boolean antecedents that makes better predictions for the negated-conjunction examples. Third, they show that disjunctions and classically equivalent negated conjunctions behave differently, thus motivating the use of Inquisitive Semantics, in which only disjunctions are inquisitive.

However, the proposal also has certain limitations. It makes incorrect predictions about certain counterfactuals with disjunctive and negated antecedents, including some negated conjunctions. In addition, the obvious extension of CZC's propositional semantics to negated indefinites and universals makes strikingly incorrect predictions in some cases.

I will suggest a fix that maintains the core of CZC's proposal, but makes use of a more elaborate way of choosing interventions on the basis of the material in the antecedent. Instead of requiring (in effect) that every way of intervening to render the antecedent true also makes the consequent true, we choose interventions probabilistically, by reasoning about how the antecedent could have come about given the information encoded in the causal model.

2 Non-classical disjunction and causal counterfactuals

CZC experimentally demonstrate a failure of intersubstitutability of classically equivalent propositions in counterfactual antecedents. Consider the scenario **Two Switches**: binary switches A and B are configured so that a light is on (L) iff both are in the same position ($A \wedge B$ or $\neg A \wedge \neg B$). Right now both are up, and the light is on ($A \wedge B \wedge L$).

(1) a. If switch A or switch B were not up, the light would be off. $[\neg A \vee \neg B > \neg L]$
b. If switch A and switch B were not both up, the light would be off. $[\neg(A \wedge B) > \neg L]$

Most experimental participants who saw (1a) judged it true, but most who saw (1b) judged it false or indeterminate. This is despite the fact that (1a) and (1b) are classically equivalent.

CZC account for these examples in two steps. First, they adopt Inquisitive Semantics [4], in which disjunctions are inquisitive but negated conjunctions are not. As [3] describes in detail, Inquisitive Semantics predicts that the default reading for conditionals with disjunctive antecedents will validate "Simplification of Disjunctive Antecedents" (SDA) ([16, 19], etc.; see [1] for a similar Alternative Semantics theory). SDA is the entailment from *If ϕ or ψ, then χ* to *If ϕ, then χ, and if ψ, then χ*. This is enough to account for the preference for "true" in (1a).

Since negated conjunctions are not inquisitive in Inquisitive Semantics, we do not expect SDA in (1b). However, the example is still problematic: if theories of counterfactuals based on minimal revision or maximal similarity were to simply go Inquisitive, they would continue to make incorrect predictions for (1b). The fact that most participants judged (1b) false or indeterminate indicates that, when reasoning about the counterfactual supposition that A and B are not both up $[\neg(A \wedge B)]$, they consider the possibility that the reason that they are not both up is that both are down $[\neg A \wedge \neg B]$. Since this configuration would result in the light still being on, participants do not endorse (1b) unreservedly. However, $\neg A \wedge \neg B$ does not correspond

to a "minimal" revision of the current scenario, which has $A \wedge B$—at least, not in any intuitive sense of "minimal". There are two more minimal revisions: either turn A off and leave B on, or turn B off and leave A on. Both of these modifications would make the antecedent true while turning the light off. So, a theory based on maximal similarity/minimal revision would seem to predict incorrectly that the possibility of $\neg A \wedge \neg B$ should be ignored, rendering (1b) true.

To deal with (1b), CZC adopt a variant of Pearl's semantics based on interventions on causal models [20]. In their model of **Two Switches** there is one causal law— L is a joint effect of A and B $[L \leftrightarrow (A \leftrightarrow B)]$. There are two contingent facts: A and B. To evaluate a counterfactual, intervene to make the antecedent true and consider what follows by causal laws, pruning facts that contribute to the falsity of the antecedent or depend causally on a fact that does. (This summary is necessarily compressed and informal; see [5] for the technical details.) The counterfactual is true iff the consequent is a logical consequence of the causal laws together with the pruned facts and the antecedent. Put another way, the consequent must be true in all models that are consistent with causal laws, antecedent, and pruned facts.

So, for example, we evaluate (1b) by removing all facts that contribute to the falsity of $\neg(A \wedge B)$—which, in this case, are A and B. As a result, the factual basis is empty. There are three kinds of models consistent with the laws. Some have $A \wedge \neg B$, rendering the consequent $\neg L$ true; some have $\neg A \wedge B$, also rendering $\neg L$ true; and some have $\neg A \wedge \neg B$, rendering $\neg L$ false. Since $\neg L$ fails to be true in all of these models, the counterfactual is not true, as desired.

This result constitutes a substantial improvement on standard theories of counterfactuals (which, absent further elaboration, make the wrong prediction for (1b)) and on Pearl's (which makes no predictions about (1a) or (1b)). However, the requirement that the consequent be true in all models that are consistent with causal laws plus pruned facts turns out to be too strong: there are cases where some of the models seem to matter more than others. I'll present the examples first, and the propose a way to make sense of them in terms of explanatory reasoning.

2.1 First puzzle: Failures of SDA.

The use of intervention makes the type of counter-examples to SDA noted by [18] especially acute for CZC. The basic observation is that, when the disjuncts vary substantially in plausibility, the counterfactual supposition may be biased toward the more plausible disjunct.

(2) If it were raining or snowing in Washington, D.C., it would be raining.

By SDA, this should imply *If it were raining in D.C., it would be snowing*, which is absurd. This has often been taken to refute SDA as a semantic principle, but the issue is subtle. Proponents of SDA have objected that the implication has inappropriate presuppositions [8], and that snow in D.C. is being treated as impossible, so that the implication is vacuously true [23, 25, 26]. However, there are related counter-examples to SDA that can't be dismissed in this way.

(3) If it were raining or snowing in D.C., it's likely, but not certain, that it would be raining.

Both of (3)'s entailments by SDA are unsatisfiable. So, SDA is not generally valid.

(4) a. If it were raining in D.C., it's likely, but not certain, that it would be raining.
b. If it were snowing in D.C., it's likely, but not certain, that it would be raining.

The fact that SDA is not always appropriate is not in itself a problem for CZC: all that is needed is an optional semantic operation that can flatten an inquisitive disjunction into a classical disjunction. When this operation is applied, a disjunctive antecedent is equivalent to a negated conjunction. So (2) should be equivalent to (5), and (3) to (6).

(5) If it weren't both not-raining and not-snowing in D.C., it would be raining.

(6) If it weren't both not-raining and not-snowing in D.C., it's likely, but not certain, that it would be raining.

All of these examples are then interpreted like (1b): we throw out facts that contribute to the falsity of the antecedent—here, ¬**rain** and ¬**snow**—and ask what holds in all consistent models. One consistent model has **rain** ∧ ¬**snow**, rendering **rain** true. Another has ¬**rain** ∧ **snow**, rendering **rain** false. Since **rain** cannot be true in all such models, (2) is necessarily false for CZC even on the interpretation that does not validate SDA. Similarly, (3) will turn out false when the theory is supplemented with a plausible treatment of epistemic operators, which should validate the obvious *If ϕ were the case then it's certain that ϕ would be the case.*

2.2 Second puzzle: Partial retention

In the famous **Firing Squad** scenario, riflemen A and B are ready to execute a prisoner. The colonel gives the order (C), and simultaneously A fires (A) and B fires (B). The prisoner dies (D). The laws implicit in the scenario are $\{C \supset A, C \supset B, (A \vee B) \supset D\}$. Now consider (7):

(7) If A and B hadn't both fired, the prisoner would still have died. $[\neg(A \wedge B) > D]$

For CZC (7) is not true, by the same logic as (1b) in **Two Switches**: one way for the riflemen not to *both* shoot is for them to both refrain from shooting. I find this result unsatisfactory, since I can readily imagine judging (7) true along the following lines: if they had not *both* fired, *one of them* would still have fired, since it's extremely unlikely that both would independently and simultaneously (e.g.) have a rifle malfunction, or decide to risk court-martial by disobeying their colonel. Admittedly, the intuition here is not totally compelling. (I will try to explain why below.) A starker issue is CZC's incorrect prediction that (8)-(9) cannot be true under any circumstances, as long as A and B are independent (given C) and both are possible.

(8) If A and B hadn't both fired, one of them would still have fired. $[\neg(A \wedge B) > (A \vee B)]$

(9) If A and B hadn't both fired, the prisoner would still have died, since they wouldn't *both* have had a rifle malfunction. $[\approx \neg(A \wedge B) > D \wedge \neg(A \wedge B) > (A \vee B)]$

Example (7) may be confounded, for example, by the interpretation of *both* and/or focus. In addition, we might rationalize the fact that A and B did not both fire by backtracking to C, considering the possibility that the colonel did not give the order (so that neither would have fired)—though this strategy would not allow us to make sense of (8)-(9). In any case, the same issues arise with other examples. (10) avoids these confounds and is readily read as being true.

(10) If the colonel had given the order and riflemen A, B, C, D, E, F, G, H, I, and J had not all fired, the prisoner would still have died.

On CZC's account we remove facts contributing to the falsity of the antecedent—A fired, B fired, etc.—and ask if the consequent follows. It does not, since there is a model consistent with the laws where the prisoner survives: the one where none of the riflemen fire.

A related example involving universal quantification makes a similar point. Imagine that we are at a Rolling Stones concert with 90,000 screaming fans. I say to you:

(11) If not all of these people had shown up tonight, there would still be a lot of people here.

This is presumably equivalent to (12), with a negated conjunction in the antecedent:

(12) If it weren't the case that (person 1 showed up and person 2 showed up and ... and person 90,000 showed up), there would still be a lot of people here.

Once we remove all facts contributing to the falsity of the antecedent—*Person i showed up* for $i \in \{1, 2, ..., 90000\}$—the consequent clearly does not follow: what if only 3, or 2, or 1, or 0 people showed up? We need an account of why these scenarios are somehow less prominent in reasoning about the counterfactual than ones that are more similar to the actual situation, where (for example) 80,000 or 89,000 show up.

2.3 Third puzzle: Indefinite and negated non-binary antecedents

I have a beagle. If I had a different kind of dog instead, I'd probably have a schnauzer, though I might have a pug. (I would never have more than one dog at the same time, though.)

This is an unremarkable kind of reasoning, but it is difficult to make sense of within interventionist theories, for two reasons. First, it is unclear what intervention is intended: there are many incompatible ways to instantiate the antecedent *If I had a different kind of dog instead*. Second, it is unclear how to make sense of the *probably . . . might . . .* in the consequent: surely, however we intervene to give me a different kind of dog, it's either a schnauzer or not. (Compare Einstein's "If I were not a physicist, I would probably be a musician" discussed in §1.)

The most comprehensive interventionist treatment of complex antecedents to date (CZC's) does not address indefinite antecedents explicitly. But there is an obvious extension: treat indefinites as disjunctions, which can be inquisitive or not. If the antecedent is inquisitive, it is equivalent (by SDA) to (13a). If it is not it is interpreted roughly as (13b).

(13) a. If I had a bulldog I'd probably have a schnauzer, but I might have a pug; and if I had a schnauzer I'd probably have a schnauzer, but I might have a pug; . . .

b. If my pet were in the set **dog** − **beagle**, I'd probably have a schnauzer, but . . .

(13a) is false, assuming I have at most one dog. For (13b), CZC require that the consequent be true in every way of making the antecedent true—i.e., no matter what kind of non-beagle dog I end up with. This cannot be true either. Even if there were only three dog breeds—beagles, schnauzers, and pugs—(14a) and (14b) could not be true (assuming ≤ 1 dog).

(14) a. If my pet were a schnauzer I'd probably have a schnauzer, but I might have a pug.

b. If my pet were a pug I'd probably have a schnauzer, but I might have a pug.

So, this example should be trivially false whether or not the indefinite antecedent is inquisitive.

This is not just a problem about indefinites. Any negation of a non-binary variable—where there are more than two possible alternatives evoked by a negated antecedent—will be associated with multiple ways to instantiate the antecedent. The most obvious extension to CZC's theory for such cases would be to require that the consequent be true under every value for the antecedent other than the one negated. Unfortunately, this won't work for negated antecedents in general. For instance, it predicts that (15a) should be true only when (15b) is.

(15) a. If I ate less chocolate I'd be thinner.

b. For any possible way of eating less chocolate, if I ate less chocolate in that way I'd be thinner.

(15a) intuitively invokes the most likely, or normal, kinds of scenarios that might play out if I ate less chocolate. (15b) is stronger: it is false unless every possible way of eating less chocolate would make me thinner. This subtle mismatch is apparent in (16), where the (a) sentence is reasonable but the attempted paraphrase in (b) is quite strange.

(16) a. If I ate less chocolate I'd probably be thinner, though I might just drink more to make up for it.

b. For any possible way of eating less chocolate, if I ate less chocolate in that way I'd probably be thinner, though I might just drink more to make up for it.

Somehow, we need to soften the interpretation of counterfactuals to focus on normal situations: requiring truth under *all* ways of intervening to make the antecedent true is too stringent.

3 Proposal: Explanatory intervention choice

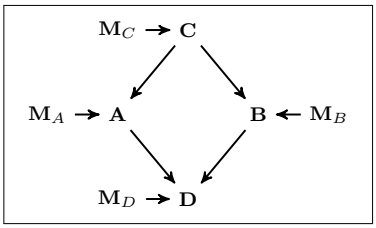
The common feature of our puzzle cases is that different ways of making the antecedent true are not even approximately matched in likelihood. Why is rain the favored instantiation of *rain or snow in D.C.*? Because rain is much more likely than snow in D.C., even though it does snow sometimes. Why, in the concert example, do we prefer to imagine a scenario where the concertgoers do not all show up by letting a smallish number staying home, rather than the entire crowd? The answer has to do with the probabilistic profile of the many, independent decisions that would be involved in the concertgoers all staying home. Since their decisions about whether to come or not are (with localized exceptions) independent, it is plausible enough that a smallish number might have decided to skip the show instead. However, it is very unlikely that a large number would have done so independently. We would have to modify a large number of independent factors to make the consequent false, thus changing the world more radically. This, I suggest, explains why (11) is so plausible.

In both cases, the diagnosis is that our background knowledge about relevant causal forces, and their probabilistic tendencies to produce scenarios compatible with the antecedent, are somehow contributing to the way that we imagine the antecedent being true. The reason that **Two Switches** is different is that the story gives us no insight into how the switches are being set. As a result, we have no basis for concluding that $\neg A \wedge \neg B$ is relatively unlikely, and the scenario where both switches are turned off is given a relatively large weight.

To model the interaction between uncertainty and the interpretation of complex antecedents formally, I will maintain the basic structure behind CZC's theory but switch to using Pearl's [20] Structural Equation Models, which incorporate an explicit representation of probabilistic uncertainty. The information in these models will be used to choose interventions for complex antecedents in a way that emphasizes *explaining* how the intervention could have come about.

To illustrate, consider a model for **Firing squad**. The laws are the same as above, but we write them as structural equations, where "=" represents assignment rather than equality. We also add for each variable V an exogenous source of randomness M_V, with prior probability $P(M_V)$, to represent uncertainty about unmodeled factors that may perturb the otherwise deterministic causal relationships represented in the model. (M is mnemonic for "malfunction".) In this example, the facts $\mathcal{F}$ are $\{C, A, B, D, \neg M_C, \neg M_A, \neg M_B, \neg M_D\}$. The box provides a graphical representation of the causal dependencies represented in the structural equation model.

- $C = \neg M_C$
- $A = C \wedge \neg M_A$
- $B = C \wedge \neg M_B$
- $D = (A \vee B) \wedge \neg M_D$



Each M_X represents a factor that could have perturbed the expected cause/effect relationship. Given that C, A, B and D are true we can infer $\neg M_C, \neg M_A, \neg M_B$, and $\neg M_D$. E.g., A's or B's rifles could have malfunctioned, each with probability p, but they did not.

The proposed procedure for evaluating a counterfactual is as follows. We first prune the facts $\mathcal{F}$ to $\mathcal{F}^*$ as in CZC, removing facts that contribute to the falsity of X or depend on a fact that does. An additional condition is needed to manage the exogenous sources of randomness: for any fact that is pruned, we also throw out the inferred values of any exogenous (M) variables that are immediately relevant to it, resetting their distribution to the prior $P(M)$.[1]

Next we consider all ways of intervening to make the antecedent true. For example, in **Firing Squad** we consider $\{\mathcal{I}_{A\wedge\neg B}, \mathcal{I}_{\neg A\wedge B}, \mathcal{I}_{\neg A\wedge\neg B}\}$, each of which would make *The riflemen do not both fire* true. Conjunctive interventions is treated as sequential intervention.

We then weight the contribution of the various possible interventions to the counterfactual. Here is one method. (There are surely further complexities in the weight function W. The weighted-intervention concept is our main positive contribution, not the precise details of this implementation.) The weight of intervention $\mathcal{I}_X$ is, up to proportionality, $W(\mathcal{I}_X) \propto P(X \mid \mathcal{F}^*)$. We combine the weights of the various possible interventions by normalization. X' ranges over the formulae characterizing the candidate interventions $\mathcal{I}_{X'}$.

$$W(\mathcal{I}_X) = \frac{P(X \mid \mathcal{F}^*)}{\sum_{X'} \; P(X' \mid \mathcal{F}^*)}$$

Normalization means that the weight of an intervention is always relative to other ways of making the antecedent true: a far-fetched possibility might receive high weight nonetheless if the alternatives are even less plausible. Note that the procedure is trivial for simple antecedents: as long as it is causally possible, the unique intervention has weight w that normalizes to $w/w = 1$.

The weight is a measure of the **explanatory value** of the candidate intervention, i.e., the extent to which it does a good job of explaining how the antecedent could have come to be true given the information encoded in the causal model. In essence, the idea is that we prefer ways of making the antecedent true that cohere with the rest of the causal model. This idea is to some extent related to explanatory backtracking (e.g. [6, 17]), but for our purposes we could get away with using backtracking only to *select among* candidate interventions.

Using the weights of the various interventions, we can find the probability of the consequent, given the counterfactual supposition in the antecedent, as the sum of the weights of the candidate interventions that make the consequent true. Some worked-out examples follow. Note that the probabilistic orientation of the proposal gives us an immediate line on the *probably* counterfactuals ((3), (13), etc.) that were troubling for SDA, for the standard interventionist semantics, and for CZC alike: we simply require that the probability assigned to the counterfactual by the method proposed above exceed the relevant threshold (see [13, 14, 27], etc.).

3.1 The Firing Squad and the Stones

In **Firing Squad** we consider *If the riflemen had not both fired, ...* To fix intuitions, let's assume that the colonel will almost certainly give the order: $P(\neg C) = P(M_C) = .01$—while rifle malfunction (willingness to risk court-martial, etc.), is slightly more likely—$P(M_{A/B}) = .1$. $\mathcal{F}$ is $\{A, B, C, D, \neg M_A, \neg M_B, \neg M_C, \neg M_D\}$. All these facts are contribute to the falsity of the antecedent, depend on a fact that does, or contribute randomness to a pruned fact; so, $\mathcal{F}^* = \emptyset$.

[1] This is a first pass. There are other ways that one could manage this issue, and more exploration of complex examples would be needed in order to choose among them.

The candidate interventions are $\mathcal{I}_{A\wedge\neg B}$, $\mathcal{I}_{\neg A\wedge B}$, and $\mathcal{I}_{\neg A\wedge\neg B}$. $W(\mathcal{I}_{\neg A\wedge B}) \propto P(\neg A \wedge B)$, which is just $P(M_A) \times P(\neg M_B)$—$.1 \times .9 = .09$. By analogous reasoning, $W(\mathcal{I}_{A\wedge\neg B}) \propto .09$. For the intervention where neither fires, $W(\mathcal{I}_{\neg A\wedge\neg B})$ is proportional to $P(\neg A \wedge \neg B)$. This is is the sum of the probability that the colonel does not give the order $[P(\neg C) = P(M_C) = .01]$, so that A and B do not fire, and the probability that he does but they fail to fire $[P(\neg M_C) \times P(M_A) \times P(M_B)]$. Using our illustrative values, this means that $W(\mathcal{I}_{\neg A\wedge\neg B}) \propto (.01{+}.99{\times}.1{\times}.1) = .0199$.

Normalizing these values, we find that on these assumptions about prior probabilities $W(\mathcal{I}_{A\wedge\neg B}) = W(\mathcal{I}_{\neg A\wedge B}) \approx .45$, while $W(\mathcal{I}_{\neg A\wedge\neg B}) \approx .1$. Since the prisoner dies in the first two interventions but not the third, the probability of example (7) (*If the riflement had not both fired, the prisoner would still have died*) is equal to $W(\mathcal{I}_{A\wedge\neg B}) + W(\mathcal{I}_{\neg A\wedge B}) \approx .9$. This may help explain the sense that (7) is highly plausible (though not totally compelling) and that its plausibility is related to the striking coincidence—Two simultaneous malfunctions!—that would be required by one of the salient ways keep the prisoner alive.

The model crucially predicts that the probability of (7) is sensitive to $P(C)$, the prior probability that the colonel would give the order. This makes sense: if we had specific knowledge about the colonel—that he must given the order no matter what, or that he is soft-hearted—it may affect our intuitions about the best explanation of *the riflemen do not both fire*. In our model it would have exactly this effect. For instance, if we hold everything the same but make the colonel a softie $[P(\neg C) = .5]$ the best explanation of the riflemen's failure to fire is that the colonel did not give the order. Accordingly, the probability of (7) decreases to about .28.

For the Rolling Stones example (11) (*If not all of these people had shown up, there would still be a lot of people here*), we have to consider a huge number of interventions, each of which removes some particular subset of the 90,000 fans. Suppose that each fan i had probability $P(M_i) = .1$ of deciding not to come, and that each chose independently. Then $W(\mathcal{I}_{\textit{Fan } i \textit{ stays home and the rest come}})$ is .1. If we remove n particular fans, that intervention receives weight $.1^n$. But, for any n, there are $\binom{90000}{n}$ ways for n fans to stay home. The distribution on weights for interventions that remove n fans from the concert is thus:

$$P(\text{If not all of these people had shown up, there would be } n \text{ fewer fans here}) \propto \binom{90000}{n} \times .1^n$$

Figure 1 depicts the distribution on number of fans removed for $P(M_i) = .1$ and $P(M_i) = .9$. Note that the y-axis is in log space: the weight differences are much greater than they appear.

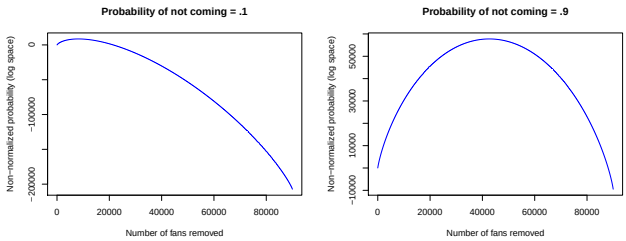


Figure 1: Weight of interventions removing n fans for *If not all of these people had shown up*

The plots show some sensitivity to $P(M_i)$, but the net effect is that there is a strong preference for interventions that remove a relatively small number of fans. The concert remains fairly well-attended ($\approx 40,000$) even if we assume that *all* of the fans were inclined to skip the Stones. The effect is a robust prediction that (11) is highly probable. This seems to be correct.

3.2 Two Switches

The key difference between **Firing Squad** and **Two Switches** is that in the latter case we know nothing about how the switches are set. Here is a simple model, where the positions of switches A and B are controlled by uncorrelated, exogenous causes:

- $A = \neg M_A$
- $B = \neg M_B$
- $L = (A \leftrightarrow B) \wedge \neg M_L$

The facts $\mathcal{F}$ are $\{A, B, L, \neg M_A, \neg M_B, \neg M_L\}$. Knowing nothing of how the switches are set, it is natural to use uninformative priors: $P(M_{A/B}) = .5$. For (1b) $[\neg(A \wedge B) > \neg L]$, all members of $\mathcal{F}$ contribute to the falsity of the antecedent, so $\mathcal{F}^* = \emptyset$. There are three interventions to consider. $W(\mathcal{I}_{A\wedge\neg B}) \propto P(A \wedge \neg B) = P(\neg M_A) \times P(M_B) = .25$. Similarly, $W(\mathcal{I}_{\neg A\wedge B}) \propto P(M_A) \times P(\neg M_B) = .25$, and $W(\mathcal{I}_{\neg A\wedge\neg B}) \propto P(M_A) \times P(M_B) = .25$.

The probability of (1b) is the normalized total weight of interventions that force $\neg L$: $(.25 + .25)/(.25 + .25 + .25) = 2/3$. This middling value may explain why so many participants chose the "Indeterminate" response option in CZC's experiment. By comparison, the disjunction (1a) should (on the SDA reading) have probability 1, and so we expect very high agreement modulo error or noise. This illustrates one way that explanatory reasoning may be able to account for the subtle intuitive differences among **Two Switches**, **Firing Squad**, and the concert example, despite their logical similarity.

3.3 Weather

To model (2)—*If it were raining or snowing in D.C., it would (probably) be raining*—recall that we have to flatten the antecedent to a classical disjunction to avoid inconsistency. Suppose that possible states of weather in D.C. are {**sun, cloud, rain, snow**}, with respective probabilities $.9, .079, .02$, and $.001$]. The only relevant fact, **sun**, is pruned, leaving $\mathcal{F}^*$ empty. Interventions that make **rain** $\vee$ **snow** true are weighted according to prior probabilities: $W(\mathcal{I}_{rain}) = .02$, $W(\mathcal{I}_{snow}) = .001$. (2) is thus true with probability $.02/(.02 + .001) \approx .95$.

4 Conclusion

Interventionist theories of counterfactuals have been hampered by the lack of a treatment of complex antecedents. CZC provide an excellent beginning, but I argued that their requirement of truth in *all* models consistent with the laws and pruned facts—is too strict. I proposed a way of using probabilistic information encoded in Structural Equation Models to weight interventions according to their explanatory value, resulting in a probabilistic interpretation of counterfactuals that maintains the core of CZC's insightful account.

The formal proposal that I have made is resolutely speculative and preliminary. In addition to exploring alternative formalizations, in ongoing work I am testing qualitative predictions regarding, for example, the way that manipulating the causal forces involved in setting the switches in **Two Switches** should influence people's responses, and quantitative predictions about exactly how probabilistic manipulations should do so. Many further questions remain, of course. In addition to the obvious linguistic connections (e.g., counterfactual donkey sentences), there are concerns about the lack of truth-conditions *per se* in the account given here. One possibility is that counterfactuals are thoroughly probabilistic, lacking truth-values (e.g., [7]). Another possibility is that truth could be defined somehow in terms of high probability. I will have to leave these questions for another time.

References

[1] Luis Alonso-Ovalle. Counterfactuals, correlatives, and disjunction. *Linguistics and Philosophy*, 32(2):207–244, 2009.

[2] Lucas Champollion, Ivano Ciardelli, and Linmin Zhang. Breaking de Morgan's law in counterfactual antecedents. In Mary Moroney, Carol-Rose Little, Jacob Collard, and Dan Burgdorf, editors, *26th Semantics and Linguistic Theory Conference (SALT 26)*, pages 304–324, Ithaca, NY, 2016. LSA and CLC Publications.

[3] Ivano Ciardelli. Lifting conditionals to inquisitive semantics. In *Semantics and Linguistic Theory*, volume 26, pages 732–752, 2016.

[4] Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. Inquisitive semantics: a new notion of meaning. *Language and Linguistics Compass*, 7(9):459–476, 2013.

[5] Ivano Ciardelli, Linmin Zhang, and Lucas Champollion. Two switches in the theory of counterfactuals: A study of truth conditionality and minimal change. *Linguistics and Philosophy*, to appear.

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

[7] Dorothy Edgington. Counterfactuals. In *Proceedings of the Aristotelian Society*, volume 108, pages 1–21, 2008.

[8] Kit Fine. Counterfactuals without possible worlds. *Journal of Philosophy*, 109(3):221–246, 2012.

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

[10] Stefan Kaufmann. *Aspects of the Meaning and Use of Conditionals*. PhD thesis, Stanford, 2001.

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

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

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

[14] Daniel Lassiter. *Graded Modality*. Oxford University Press, 2017.

[15] David Lewis. *Counterfactuals*. Harvard University Press, 1973.

[16] Barry Loewer. Counterfactuals with disjunctive antecedents. *The Journal of Philosophy*, 73(16):531–537, 1976.

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

[18] Thomas McKay and Peter Van Inwagen. Counterfactuals with disjunctive antecedents. *Philosophical studies*, 31(5):353–356, 1977.

[19] Donald Nute. Counterfactuals and the similarity of worlds. *Journal of Philosophy*, 72(21):773–778, 1975.

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

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

[22] Katrin Schulz. "If youd wiggled A, then B wouldve changed": Causality and counterfactual conditionals. *Synthese*, 179(2):239–251, 2011.

[23] William B Starr. A uniform theory of conditionals. *Journal of Philosophical Logic*, 43(6):1019–1064, 2014.

[24] Frank Veltman. *Logics for conditionals*. PhD thesis, University of Amsterdam, 1985.

[25] Ken Warmbrōd. Counterfactuals and substitution of equivalent antecedents. *Journal of Philosophical Logic*, 10(2):267–289, 1981.

[26] Malte Willer. Simplifying with free choice. 2017. In press at *Topoi*.

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