Probabilities of Counterfactuals and Counterfactual Probabilities

Alan Hájek

[NEEDS POLISHING; COMMENTS WELCOME]

1. Introduction

My topic is the interaction of probabilities and counterfactuals. Can probabilities illuminate counterfactuals: their logic, their truth conditions (or lack thereof), and their approximate-truth conditions?

It is natural to think that probabilities of conditionals are parasitic on their truth conditions. After all, that is the model that we find with the basic connectives: probabilities of conjunctions, disjunctions, and negations are parasitic on their respective truth conditions. Similarly, probabilities of quantified statements are determined by their truth conditions, rather than the other way round. And so it goes with almost all of our language that is probability-apt: it’s truth first, probability second. But a striking theme in much of the conditionals literature is that this order of dependence is reversed for conditionals: their probabilities are regarded as primary, and consequences for their truth conditions (or lack thereof) are then drawn. Offhand, then, I find it surprising that conditionals should have special status in having their semantics underpinned probabilistically. But if they should, then that fact is striking in itself.

In this paper I will discuss two notable ‘probabilities first’ accounts of counterfactuals: those presented in agenda-setting papers by Edgington (2008) and by

---

1 For valuable discussion and inspiration, I am especially grateful to Sharon Berry, Rachael Briggs, Yoaav Isaacs, Hanti Lin, Daniel Nolan, Paolo Santorio, Wolfgang Schwarz, Julia Staffel, and Hlynur Steffanson. Thanks also to David Etlin for helpful e-mail correspondence.
Leitgeb (2012 a and b). I will end with a brief manifesto of my own views on counterfactuals.

2. Edgington’s account

2.1 Outline of the account

Edgington begins with a partial job description for counterfactuals:

they have figured … in accounts of causation, perception, knowledge, rational decision, action, explanation, and so on. And outside philosophy, in ordinary life, counterfactual judgements play many important roles, for instance in inferences to factual conclusions … (1)

She has long been an advocate of a suppositional account of indicative conditionals.

On this view, a conditional statement is not a categorical assertion of a proposition, true or false as the case may be; it is rather a statement of the consequent under the supposition of the antecedent. A conditional belief is not a categorical belief that something is the case; it is belief in the consequent in the context of a supposition of the antecedent. (2)

Enter probability theory. Edgington regards the strongest evidence for the suppositional account to come from the way uncertain conditional judgments work. Our best theory of uncertainty is probability theory, and our best understanding of conditional uncertainty is conditional probability. Putting these ideas together, a core part of her account is Adams’ Thesis that the probability of an indicative conditional ‘if A, then B’ is the conditional probability of B, given A (where this is defined). But this probability here is not to be understood as that of the conditional’s truth. On the contrary, it is equally central to the account that indicative conditionals do not have truth values.

Various authors (including Adams, Bennett, Bradley, Gibbard, and McGee) subscribe to this theory of indicative conditionals alongside Edgington. But most authors believe that counterfactuals must be treated differently—typically with some
sort of ‘similarity’ semantics, à la Stalnaker (1968) and Lewis (1973). Edgington insists, however, that counterfactuals and indicative conditionals should be given parallel treatment. She argues that “the easy transition [sic.] between ‘suppose’ and ‘if’ is as evident for subjunctives as it is for indicatives.” (4-5). And it certainly seems to be a welcome consequence of her view that the word ‘if’ is given parallel treatment—especially so when indicatives and counterfactuals seem to coincide in the future tense. Above all, according to her they should be given parallel probabilistic, rather than truth conditional, treatment.

However, their treatment had better not be too parallel: witness our divergent assessments of

- if Oswald didn’t shoot Kennedy, then someone else did

and

- if Oswald hadn’t shot Kennedy, then someone else would have.

(The former seems highly assertable, while the latter does not.) According to Edgington, while indicative conditionals such as the former are governed by Adams’ Thesis, counterfactuals such as the latter go by a different conditional probability:

…conditional probability— the probability of C on the sup-position [sic.] that A—is, I claim, the key to how close to certain we are of a conditional of any kind. For subjunctive conditionals, this conditional probability does not (normally) represent your current degree of belief in C given A, but (most typically) your view about how likely it was that C would have happened, given that A had” (5)

The locution “your view about how likely it was …” has two components: “your view”, and “how likely it was”. This combines a subjective and an objective element. It is clear from the preceding material that Edgington regards the objective element to be conditional chance at the appropriate time. The subjective element needs some massaging, as you may not have a single “view” as to that conditional chance. Rather,
you may spread credences over various hypotheses about what it is. I think it is best to think of this as in some sense your expectation of the conditional chance\(^2\): the weighted average of the conditional chance according to each hypothesis, the weights being in some sense your credences for each hypothesis that it is the correct one.

This is very much along the lines of Skyrms’ (xx) and Leitgeb’s (2012) measure of the acceptability of a counterfactual. However, more needs to be said about the sense of credences that is operative here. Edgington writes: “the conditional probabilities relevant to the assessment of subjunctive conditionals do not (typically) represent your present actual distribution of belief, but those of a hypothetical belief state in a different context” (8). Which hypothetical belief state? Which context? Edgington is forthright about this posing a problem:

suppose such-and-such had been the case; what do you hold constant? For Goodman this is the problem of cotenability, for Lewis it’s the problem of closeness, for me, it’s the problem of which context shift is appropriate, which probability distribution is the appropriate one, given that it is not the one that represents your present state of belief. (13-14).

Later I will suggest a solution to this problem.

So the central pillars of Edgington’s account are that counterfactuals have probabilities (governed by a Skyrms-like connection to conditional chances), but they do not have truth values. I now want to put pressure on both of these pillars, and in particular, their combination.

### 2.2 Some questions about Edgington’s account

I want to raise a number of questions about Edgington’s account. In the following section I will suggest some answers, though I am not sure that she would endorse them.

\(^2\) Thanks here to Rachael Briggs.
i) What does it mean for Edgington’s “probabilities” of counterfactuals to obey the probability calculus?

Here and elsewhere I put the word “probabilities” in scare quotes, because for a number of reasons I am not sure that they really are probabilities, or that they should be. Indeed, it is not even clear what it would mean for them to obey the probability calculus. What sense are we to make of the additivity rule as applied to them? It is an axiom of probability theory that the probability of a disjunction of incompatible sentences is the sum of their individual probabilities. There are two problems here for truth-valueless sentences: disjunction, and incompatibility.

First, disjunction. It is a truth-functional operation; how should it be defined for sentences that lack truth values? To be sure, Edgington’s account is solely about single counterfactuals, not more complex sentences in which they might appear, such as disjunctions. Yet such disjunctions may have perfectly determinate meaning. For example, I regard the following to be a tautology:

‘if Oswald hadn’t shot Kennedy, someone else would have or

it’s not the case that if Oswald hadn’t shot Kennedy, someone else would have’

Second, incompatibility. This is usually defined in terms of truth: two sentences are incompatible if it is not possible for both of the sentences to be true. Either this notion makes no sense for truth-valueless sentences, or it trivially applies to all pairs of such sentences (for one way to fail to be true is to lack a truth value).

I say that it follows from the probability calculus that

\[ P(\text{if Oswald hadn’t shot Kennedy, someone else would have}) + \\
P(\text{it’s not the case that if Oswald hadn’t shot Kennedy, someone else would have}) = 1 \]
But can Edgington say this? It follows if the two ‘Oswald’ sentences form a partition, picking out mutually exclusive and jointly exhaustive propositions. However, how do sentences that lack truth value form a partition? In failing to express propositions, they fail to exclude or to exhaust anything.

Now, it may seem obvious how we should understand ‘incompatibility’ for truth-valueless but probability-bearing sentences: not in terms of truth, but in terms of probability! For example, we might say that $p$ and $q$ are probabilistically incompatible just in case $P(p \text{ and } q) = 0$. And it may seem obvious that Edgington should understand ‘incompatibility’ for counterfactuals along these lines.

I have three concerns with this proposal.

**Concern 1.** Probabilistic incompatibility is a poor surrogate for good old-fashioned logical incompatibility. The latter is two-place relation between a pair of sentences; but the former is a three-place relation between a pair of sentences and a probability function. Change the probability function appropriately, and you will change a verdict of probabilistic compatibility (trivial cases aside). Moreover, there are various irregular probability functions, which assign probability 0 to contingent propositions. For example, suppose that we toss a fair coin infinitely many times. The probability that it lands heads each time is 0, according to the relevant probability function. (Williamson 2007). Then the sentence ‘the coin lands heads each time’ is probabilistically incompatible with itself! Yet far from being logically incompatible with itself, it is equivalent to itself—while being contingent.

**Concern 2.** Probabilistic incompatibility is defined in terms of probability. But what is probability? Well, it’s characterized by the axioms of probability: additivity, … Hang on—the problem was that it is unclear how to make sense of the additivity axiom for truth-valueless sentences, and thus unclear how to make sense of probability
assignments to them; it is no response simply to appeal to probability assignments to them!

**Concern 3.** The definition I have suggested of probabilistic incompatibility requires a *conjunction* of sentences (‘p and q’). However, Edgington’s account applies to single counterfactuals; I am not sure how it applies to compounds such as the conjunction of a counterfactual with another sentence. Again, conjunction is a truth-functional operation; how does it operate on sentences that lack truth values?

Perhaps the supervaluational approach to vagueness gives some hope for her proposal? We want to say that the sentences ‘Fred is bald’ and ‘Fred is not bald’ form a partition, even if it is indeterminate whether Fred is bald. The supervaluationist can say this, because however we precisify ‘bald’ (say, according to number of hairs), the two sentences are mutually exclusive and jointly exhaustive. And since ‘Fred is bald or Fred is not bald’ is supertrue, its probability is 1. However, a supervaluational approach to conditionals is surely not in the Edgingtonian spirit. Indeed, she explicitly rejects Stalnaker’s supervaluational approach (p. 14). Stalnaker begins with his (1968) truth conditions for ‘if it were that p, it would be that q’: q is true at the closest p-world. Now suppose there is more than one world tied for the closest p-world. The counterfactual is true if q is true at all of these worlds, and it is *indeterminate* if q is true at some of these worlds but not others. It is no part of Edgington’s ‘no truth value’ proposal that conditionals have indeterminate truth value.

I have asked what it even means for Edgington’s counterfactuals to have probabilities. However, let us suppose for the sake of the argument that we can adequately answer this question. A further question then immediately arises.
ii) Why should Edgington’s “probabilities” of counterfactuals obey the probability calculus?

In order to deserve the name, Edgington’s “probabilities” of counterfactuals must obey the probability calculus. But why should they? There are various arguments that credences assigned to sentences should do so, on pain of irrationality; but most of these arguments are parasitic on the sentences having truth values.

Consider the Dutch Book argument for probabilism. We identify your credences with your betting prices, and show that if you these are not probabilities, then you are susceptible to a Dutch Book: a set of bets each of which you will regard as acceptable, but which collectively guarantee your loss. However, it is hard to make sense of betting on something that has no truth value: there are no conditions for determining whether the bet wins or loses. Note that this is not merely the old concern about the betting interpretation that it misrepresents one’s attitudes to unverifiable propositions, which can never be settled favourably. There, at least, it is clear what it takes for a bet on such a proposition $p$ to win: it does so just in case $p$ is true. Whether or not anyone realizes or verifies that it is a winning bet is another matter. We might imagine God settling the matter. But if $p$ has no truth value, there is nothing for God to settle: there are no winning conditions for the bet. If we can make sense at all of such a bet, it is clear what its fair price is: zero. After all, it is a bet that cannot pay off. In sum: either we can make no sense of betting on counterfactuals on the Edgington account, or we can, and their betting prices should always be 0. Either way, the Dutch Book argument fails.

Or consider the calibration argument for probabilism. Calibration is a measure of how well one’s credences match the corresponding relative frequencies. For example, suppose that each evening you assign a credence to it raining the following day. We
keep track of each assignment and of whether it rained or not the following day. You are perfectly calibrated if for each $p$, among those days for which you assigned credence $p$, proportion $p$ of them were rainy—proportion $p$ of the sentences stating that the corresponding days were rainy are true. More generally, we can measure how well calibrated your assignments are with respect to a set of sentences, where this comes in degrees that may fall short of perfection. We can then show that if your assignments violate the laws of probability, then there is a probability function whose assignments are guaranteed to better calibrated than yours. However, it is difficult to make sense of calibration with respect to a set of sentences that don’t all have truth values. Edgington cannot appeal to the calibration argument in support of her probabilities of counterfactuals obeying probability theory, for by her lights there is nothing to which these probabilities can be calibrated.

Or consider the accuracy argument for probabilism. Every time you assign a credence to a proposition, there is a fact of the matter of the distance by which your credence missed the proposition’s truth value (identifying 1 with truth and 0 with falsehood). The smaller this distance, the more accurate your credence is. We may then give an overall accuracy score to your entire credence function, according to how well it fares accuracy-wise over all the propositions in its domain. We can then show that if your credence function is not a probability function, then there is a probability function that is guaranteed to score better than your function. However, it is especially clear that no sense can be made of ‘accuracy’ with respect to sentences that lack truth values, and thus that do not express propositions. Edgington cannot appeal to the accuracy argument in support of her probabilities of counterfactuals obeying probability theory, for by her lights they have no truth values by which accuracy can be defined.
So some of the leading arguments for probabilism fail for assignments to counterfactuals on the Edgington account. That said, perhaps one class of such arguments can be deployed: those based on representation theorems. We lay down some rationality axioms on your (qualitative) preferences, and show that if you obey these axioms, then you are representable as maximizing expected utility according to a utility function and a probability function. So if your preferences are not representable with a probability function, you must violate some of the axioms, and hence be irrational. Interestingly, this argument need not have any particular connection to truth. To be sure, on some formulations of the relevant representation theorem it will. For example, the Bolker-Jeffrey representation theorem takes the objects of preference to be propositions, and thus bearers of truth value. However, we might also think of preferences as attaching to other kinds of objects. And perhaps these objects include truth-valueless counterfactuals. Bradley (1998) offers such a representation theorem for preferences that include such counterfactuals.

Let’s grant that it makes sense to speak of Edgington’s “probabilities” of counterfactuals obeying the probability calculus, and that they should. Still, there is the further question of understanding what this means.

### iii) What do Edgington’s “probabilities” of counterfactuals mean?

As a slogan, I submit that probability is probability of truth. When I assign probability 1/6 to the sentence “this die lands 6 on the next toss”, the 1/6 attaches to the truth of this sentence (as opposed to its falsehood, or its lacking a truth value, or its grammaticality, or succinctness, or some other property of the sentence). What then, is the “probability” of a counterfactual, if not the probability of its truth? If $X$ is not truth-apt, then nor is it probability-apt. How are we to understand a locution like ‘$X$ is
probable, but it is not probably true’? Indeed, Edgington is committed to odd-sounding claims of the form: ‘if A were the case then B would be the case’ is probable, but it is guaranteed that it is not true’.

Edgington criticizes Stalnaker’s supervaluational approach on the grounds that it will accord to some counterfactuals probabilities that are too low. She imagines, for example, a doctor who thinks it is 90% likely that I would be cured if I had a particular operation. There are some closest operation-worlds at which I am not cured, and so this counterfactual turns out to be “definitely not true” by Stalnaker’s truth condition (14). Moreover, the doctor is certain that this truth condition does not obtain, so she presumably should give the counterfactual probability 0—not the value of 90% that we wanted. Ironically, however, this argument appears to backfire on Edgington: by her lights, the counterfactual is also “definitely not true”—it has no truth value—and we may assume that the doctor is certain of this. So by parallel reasoning, she should give the counterfactual probability 0 again. Of course, Edgington will insist that the probability of the counterfactual is not the probability of its truth; but then she should allow Stalnaker the same escape route.

Probability is a kind of modality (a degreed modality). Consider other modal locutions, such as ‘necessarily, p’, ‘possibly, p’, and ‘it is impossible that p’. Now append to each of these the words ‘… but p has no truth value’. I cannot make sense of the resulting sentences. What is it about p that is necessary, or possible, or impossible, if not its truth? Nor can I make sense of attaching a probability to p, but then appending those words. What is about p that has this probability, if not its truth?

However, let us suppose for the sake of the argument that it makes sense to speak of probabilities of counterfactuals without these being probabilities of their truth. That still makes counterfactuals odd critters.
iv) Are conditionals are loners on the Edgington account?

Are counterfactuals and indicative conditionals to be the only sentences of our language that have probabilities, but not truth values? We are familiar with other sentences that lack truth values:

- imperatives, such as “Shut up!”
- questions, such as “Who is that in the gorilla suit?”
- optatives, such as “May your dress sense improve!”
- hortatives, such as “Go, Collingwood!”
- exclamations, such as “Yay!” (after a Collingwood win)
- expletives, such as “xxxx!” (after a Collingwood loss).

But they also lack probabilities, and probability-related locutions crash on them:

- “The probability of ‘shut up!’ is 0.3”
- “The chance of ‘who is that in the gorilla suit?’ is 0.8”
- “The conditional probability of ‘may your dress sense improve!’, given ‘go, Collingwood!’, is high.
- ‘xxxx!’ is three times as probable as ‘yay!’ (Though admittedly I have been to faculty meetings where I was somewhat tempted to say something like that.)

And this pattern seems to be no accident. Conditionals, on the Edgington view, thus seem to live in a strange halfway house: lacking probabilities (like all these kinds of sentences), but having probabilities (unlike any of them).

That said, she is not alone in giving conditionals special treatment—recall my introductory observation that much of the conditionals literature reverses for them the usual order of dependence of probabilities on truth. And perhaps conditionals are not
entirely alone in that house after all. Perhaps epistemic modals live there too. Consider ‘the butler might have done it’, said in a context in which I am not sure whether the evidence conclusively rules out his guilt or not.\(^3\) Perhaps I may assign this intermediate probability, while on various accounts of the epistemic modal ‘might’, it lacks a truth value. Perhaps moral claims, and claims about what is good, and what is rational, also live there on a Gibbardian expressivist account of them. If so, some of the puzzlement that I have voiced over the last few pages carries over to these other putative inhabitants of the halfway house. But at least they are not alone.

2.3. How we might answer these questions

I have raised several questions about Edgington’s account, focusing on what I take to be its unstable combination of according counterfactuals probabilities, but not truth values. Here I want to suggest how we might answer them.

I have recommended this gloss on Edgington’s account: your probability of the counterfactual ‘if it were that A, it would be that B’ is in some sense your expectation of the conditional chance of B, given A. Now let me say more about what that sense might be.

It is clear that the expectation is not calculated according to your current credences. That might be appropriate for judging indicative conditionals, but not counterfactuals. We want an appropriate hypothetical credence. Moreover, in keeping with Edgington’s suppositional account, it should be the credence appropriate under a supposition. Now, we may suppose both indicatively and subjunctively. The former involves supposing something to be actually the case, and then reasoning under that supposition. The latter (typically) involves supposing something taken to be non-

\(^3\) Thanks to Daniel Nolan for the example, and for the cases that immediately follow.
actual, but reasoning about what *would* be the case under that supposition.

*Imaging* is a good candidate for representing subjunctive supposition. Start with probability function $P$. Among other things, $P$ assigns probabilities to individual worlds. Define a new probability function $P_A$ as follows: for each world $w$, $P_A$ shifts the probability that $P$ assigns to $w$ to the closest $A$-world to $w$. $P_A$ is said to be derived from $P$ by *imaging on* $A$. Now let $P_A$ be the probability function according to which the expectation of the conditional chance is calculated.

Let ‘$A \Box \rightarrow B$’ denote the counterfactual ‘if it were that $A$, it would be that $B$’, and ‘$Cr$’ denote a rational agent’s credence function, representing her degree of confidence for each world $w$ that that world is actual. Each world $w$ has an associated chance function $ch_w$. Our agent knows for each world what this function is, but since she is uncertain about which world is actual, she is uncertain about which chance function is operative in her world. Here, then, at a first pass is the proposal presented more formally:

$$Cr(A \Box \rightarrow B) = \Sigma_w Cr_A(w) ch_w(B \mid A),$$

where $Cr_A$ is the result of imaging $Cr$ on $A$.

As far as I can tell, this allows $A \Box \rightarrow B$ to have a truth value. I know of no triviality result against *this* proposal (although whenever I see probabilities of conditionals on one side of an equation and conditional probabilities on the other, I start getting jittery!). So let me assume that counterfactuals express propositions. Then it makes perfect sense to speak of credences of counterfactuals obeying the probability calculus. Moreover, they *should* do so, since the right hand should do so. Conditional chances obey the probability calculus; credences over worlds should do so; hence credence-weighted expectations of conditional chances should do so. Counterfactuals have both truth values and probabilities on this proposal, so they are not loners.
There is a complication. Edgington realizes that her account needs some tweaking because of cases like this:

the car breaks down on the way to the airport and I miss my flight. ‘If I had caught that plane I’d be half way to Paris by now’, I remark to the mechanic, who has just been listening to the radio. ‘You’re wrong’, he tells me. ‘It crashed. If you had caught that plane you would be dead by now.’ …

This example and others like it suggest that the conditional probability we are concerned to estimate, for counterfactuals (and in a sense the ultimate verdict on some forward-looking wills), is the chance, at a time when A still had some chance of coming about, of C given A and any relevant, causally independent, subsequent facts that have a causal bearing on C. (15)

This suggests the corresponding tweak to my proposal above:

\[ Cr(A \square \rightarrow B) = \sum_w Cr_A(w) \cdot ch_w(B \mid A \& I), \]

where \( I \) is the conjunction of any facts, subsequent to \( A \), that are causally independent of \( A \), but that cause \( B \) (either directly, or by being causal ancestors of \( B \)).

The thought is something like this. The plane in fact crashed, and something caused it to do so. We are to imagine that whatever this cause was, it was independent of Edgington catching the flight. We should then conjoin the fact that specifies this cause to the antecedent to form the condition of the conditional expectation; we then take the expectation of that quantity, much as before.

There is something to this idea, but I think it isn’t quite right as it stands. Let us suppose that the plane crashed because it was overloaded with people, and it was as a result too heavy. Now suppose that Edgington had caught the plane. It would have been even heavier (only slightly, to be sure!). Then, all the more the plane would have crashed. We want to factor in the crashing of the plane, even though it is not causally independent of the antecedent.

[Or do we want to say that it is causally independent of the antecedent: Edgington’s
presence on the plane would have been causally idle, since the crash was bound to happen either way? Then we should say that everyone’s presence on the plane would have been causally idle. (I am assuming that no single person was the causal difference-maker.) And yet the weight of the passengers taken collectively caused the plane to crash. I’m not sure about this!]

There is also a problem that antecedents of counterfactuals are not always datable, and so there may not be any facts subsequent to them. Consider ‘if gravity obeyed an inverse cube law, the planets would not follow elliptical orbits’. Or continuing one of Edgington’s own examples (footnote 2, p. 1), ‘if Euclidean geometry had been true, triangles in physical space would have had angles summing to exactly 180 degrees’. There is no time at which we are supposing that suddenly gravity started obeying an inverse cube law, or that suddenly Euclidean geometry became true. Rather, we are supposing that gravity always did obey the law, and that Euclidean geometry always was true.

Be all that as it may, I offer my tweaked proposal as a friendly suggestion as to how Edgington’s position should be understood. However, I doubt that she will welcome it, as it rejects an essential part of that position: that counterfactuals do not have truth values.

Let me turn, then, to a related account of counterfactuals that does accord them truth values—that of Leitgeb. We will see that while he solves some of the problems that I have raised for Edgington, he still faces some others.

3. Leitgeb’s account

3.1 Outline of the account

In his rich papers (2012 a and b), Leitgeb offers a new semantics for
counterfactuals. Like Edgington’s, it is probabilistic; unlike Edgington’s, it accords truth values to counterfactuals. An important motivation of his is naturalistic: to “ground counterfactuals semantically in a scientifically respectable manner” (26). Relatedly, he hopes his probabilistic semantics will “state what the correct truth conditions for many occurrences of counterfactual statements in natural language ought to be like.” (60) What he offers is thus more a Carnap-style explication than a conceptual analysis of counterfactuals as we find them in the wild. I can hardly do justice to his two papers in this short space. For example, I will not touch on his two kinds of pragmatic meanings of counterfactuals, his triviality result, or his representation theorems. Instead, I will concentrate on his truth conditions.

I take there to be three main steps leading to them. They proceed as follows. Step one:

[consider] the subjunctive conditional

(1) If the match were struck, it would light
… As step one, it should be unproblematic to qualify the consequent of (1) in either of the following ways, while leaving the meaning of the conditional—completely or at least almost completely—unaffected:

(2) If the match were struck, it would necessarily [definitely] light.
(3) If the match were struck, it would be very likely to light.

Without any bias from some philosophical theory, (2) and (3) simply seem to say pretty much the same as (1). (36)

Step two:

Next, we intend to analyze (3) with the aim of reformulating it somewhat more transparently. A sentence such as (3) invites two obvious questions: Which kind of probability is expressed by the qualification ‘very likely’? And what is the logical form of the whole conditional statement? … The probability in question is an ontic one—single-case chance …—and the logical form of (3) is the one of:

(4) The conditional chance of the match lighting given that it is struck is very high. (37)…

Generalizing from this particular case of Leitgeb’s, I take it the main steps so far are:
Step 1. ‘if $p$ were the case, $q$ would be the case’ means almost the same as ‘if $p$ were the case, $q$ would very probably be the case’.

Step 2. The latter is analysed as: $ch(q \mid p)$ is very high, where $ch$ is the objective chance function (suitably relativized to a world and time). “Very high” is understood as “above a vaguely determined threshold” (113).

The third and final step is to give a precise mathematical characterization of such conditional chances. He wants to allow for them to be defined even when the condition $p$ has chance 0. This can be achieved by taking conditional chances as primitive, and there is a framework to represent this:

Step 3. $P$ is a Popper function.

In a nutshell, this is a version of what Bennett (2003) calls the “near-miss proposal”:

$$p \square \rightarrow q \text{ is true iff } ch(q \mid p) \text{ is very high.}$$

Again, I elide over many details (e.g. the time to which the chance function should be indexed); but I think that steps 1 – 3 capture the essential features of Leitgeb’s account. The appeal to chance secures the objectivity of the truth conditions for counterfactuals; their conditionality is captured by the appeal to conditional chance.

I will now discuss these steps in turn.

3.2 Questioning Leitgeb’s account

Questioning Step 1

There may be close connections between the truth of a proposition and its having very high chance, but I do not think that near synonymy is one of them. “It will rain tomorrow” and “it will very probably rain tomorrow” usually have the same truth value, and believing one will typically have the same impact on one’s behaviour as
believing the other (though not always). However, their meanings are quite different. A sunny, rain-free day tomorrow falsifies the first, but not the second.

And so it is with counterfactuals, I would have thought. There is something defective about

“If the match were struck, it would light; but if the match were struck, it might not light.”

Following Lewis’s (1973), I think it is the defect of semantic inconsistency. DeRose thinks it is that of pragmatic inconsistency, but that is still defect enough for him to call the tension between ‘would’ and ‘might not’ an “inescapable clash”. But there is nothing defective whatsoever about

“If the match were struck, it would be very likely to light; but if the match were struck, it might not light.”

There is no clash at all between ‘would very probably’ and ‘might not’. On the contrary, ‘would very probably’ seems to conversationally implicate ‘might not’. Our inner Gricean infers that if ‘… would very probably’ is the strongest thing that can be said, ‘… might not’ must be the case! Leitgeb notes that “it is clear that might-conditionals should not be represented as the duals of the would-conditionals that obey our semantics” (111). He surmises that at best, one may turn to ‘would’ counterfactuals with possibility operators in their consequents. That is, ‘if it were the p, it might be that q’ has the form:

\[ p \Box \Rightarrow \Diamond q \]

But this seems to be a cost of his view, since again there is no clash at all between ‘would’ and ‘would be possible that not’, much as there is no clash at all between the claim that I am a philosopher and the claim that it is possible for me not to be a philosopher (at least for a non-epistemic sense of possibility).
Let us return to the putative near-synonymy of ‘would’ and ‘would very probably’. Suppose a doctor truthfully tells your father: ‘if you were to take this pill, you would very probably survive’; your father takes the pill and dies. Now suppose instead that the doctor had said: ‘if you were to take this pill, you would survive’; your father takes the pill and dies. Your case for a law suit is much stronger in the latter case than in the former. This also shows how Leitgeb’s semantics for counterfactuals allow modus ponens failures, which trouble me more than they appear to trouble him. ‘Would’ and ‘would very probably’ do not have almost the same meanings, because they license quite different inferences.

But even Leitgeb himself speaks of (3) being “… “almost” synonymous with (1)” (37), with tell-tale scare quotes around “almost”, as if he does not really mean “almost”. (Perhaps “almost almost”?) Indeed, by parallel reasoning,

‘if p were the case, q would very probably be the case’ presumably means “almost” the same as

‘if p were the case, it would very probably be the case that q would very probably be the case’.

And so on ad infinitum. I suggest that “almost” synonymy is an unreliable guide to truth conditions. This point will be important again later, when we come to his account of the “approximate truth” of counterfactuals. However, recalling that Leitgeb’s project is one of explication rather than conceptual analysis, perhaps near enough is good enough for his purposes.

**Questioning Steps 2 and 3**

Is ‘if p were the case, q would very probably be the case’ to be analysed as $\text{chance}(q \mid p)$ is very high? Here are some reasons to think not.
Conditional probability is indicative rather than subjunctive in character. This is most easily seen for conditional credence, but I think the main idea carries over to conditional chance. Credences assigned to worlds represent an agent’s uncertainty about which world is actually the case; conditional credences are concentrated on more specific propositions. Conditionalising on such a proposition amounts to learning that it is actually the case. Speaking of the chance function ‘learning’ something about what is actually the case is metaphorical, but it is a helpful way to think of conditional chances. Counterfactual reasoning, by contrast, typically involves supposing something about a hypothetical scenario, one that is taken to be non-actual. To capture the truth conditions for counterfactuals, hypothetical conditional chances would be required, those operative in some other world; not actual conditional chances, those operative in our world.

On the usual view about chance, all false propositions about the past have chance 0. For example, Oswald’s not killing Kennedy has chance 0 now (let us assume). So on Leitgeb’s proposal, many counterfactuals will correspond to conditional chances for which the condition has chance 0. To be sure, Popper functions provide a mathematical framework for which such conditional chances may nonetheless be defined. But they may also be undefined. The Popper formalism comes with no guarantee that its conditional probability functions are endowed with a rich domain, still less a complete domain of all propositions that might appear as antecedents of counterfactuals. Moreover, whether or not the conditional chance function (at a time) is well defined for this or that condition is a metaphysical question, not one that can be settled by appealing to the axioms and theorems of the Popper calculus.

Indeed, there are good reasons for thinking that such conditional chances will not be defined for various counterfactuals. Consider counterfactuals about the chances
themselves being different from what they actually are—e.g. ‘if this (actually fair) coin had chance 1 of landing heads, the coin would repeatedly land heads’. It is quite unclear that the conditional chance of the consequent, given the antecedent, is well defined. Similarly, some counterfactuals do not explicitly mention chances, but tacitly involve them nonetheless—e.g. ‘if the half-life of uranium were much longer, we would see more uranium around us’, and perhaps even ‘if the vase were less fragile, it would survive greater impacts than it actually does’.

And then there are counterfactuals for which chances just don’t seem to be germane at all—for example, the very first counterfactual that Leitgeb discusses in his paper, ‘if Oswald hadn’t killed Kennedy, someone else would have’. I have no idea whether there is a corresponding conditional chance for this counterfactual, and even if there is, I question whether its value underwrites the truth condition for the counterfactual. Consider ‘if gravity obeyed an inverse cube law, the planets would not follow elliptical orbits’. I see no reason to think that the corresponding conditional chance is high, or even defined.

While I laud Leitgeb’s goal of naturalizing counterfactuals, I wonder whether he achieves it. I am not convinced that science supports his appeal to Popper functions, or their delivering conditional chances that align with the truth values of counterfactuals, in accordance with his semantics. Somewhat curiously, none of his examples of counterfactuals are drawn from science.

3.3 Truth, or approximate truth?

To repeat, Leitgeb offers the truth condition for a counterfactual as the corresponding conditional chance being very high. It is important for me to emphasize this.
In order for [‘if I had dropped the plate, it would have fallen to the floor’] to be true according to our probabilistic truth condition, it suffices for there to be a very high conditional probability of the plate falling to the floor if dropped. But that conditional probability is in fact high—exceptions to it having a probability close to 0—even by the lights of quantum physics. So the counterfactual comes out true in our semantics according to our ... interpretation of ‘\(ch(B|A) = 1\)’ in terms of ‘the probability of \(B\) given \(A\) is very high, or close to 1’... (111, my bold. I replace his notation for the conditional chance with my own.)

This exposes his account to the objections I raised in the Questioning Step 1 section. However, he tantalizes us with what looks like a hard-line truth condition that requires the conditional chance to be exactly 1: \(ch(B|A) = 1\). I confess that I am puzzled by his interpretation of this as ‘the probability of \(B\) given \(A\) is very high, or close to 1’. To be sure, he justifies it with some formidable mathematical machinery (involving limits of infinite sequences of Popper functions). But for all that, ‘\(ch(B|A) = 1\)’ denotes an identity, not an approximate identity. Be that as it may, I think the best prospect for a probabilistic analysis of counterfactuals is the hard-line one.

But even the hard-line truth condition is not hard enough for my liking. Irregularity is again the sticking point. Probability 1 is insufficient for truth—remember the infinitely tossed coin landing heads every time. Similarly, a conditional probability of 1 is insufficient for the truth of the corresponding counterfactual, I say. The conditional probability that the coin does not land all heads, given that it is tossed infinitely many times, is 1. But I claim that the counterfactual

the coin is tossed infinitely many times \(\square\)→ coin does not land all heads

is false. For each possible infinite sequence of heads and tails, the coin might land that way; in particular, it might land all heads. This is inconsistent with the claim that it would not do so. Or so I believe, siding with Lewis’s treatment of ‘might’ counterfactuals.
‘Might’ counterfactuals, and similarly ‘might not’ counterfactuals, are easily made true: the chances of their consequents, given their antecedents, do not even need to be positive. Correspondingly, I submit, ‘would’ counterfactuals are easily made false. This is one of several arguments of mine that most counterfactuals are false. I believe that Leitgeb would do better to explicitly endorse the hard-line truth condition, bringing him closer to my position. However, I admit that it is a cost of my position that it flies in the face of commonsense. For example, ‘If I had dropped the plate, it would have fallen to the floor’ is intuitively true, while I maintain that it is false.

3.4 Lotteries and plates

Leitgeb seeks to defend commonsense. Since the conditional chance of the plate falling to the floor, given I drop it, is very high, the counterfactual comes out true according to his truth condition. And so it goes with most ordinary counterfactuals that we take to be true. However, he seems to equivocate on this when faced with another of my arguments. However high we set the bar for ‘very high’ (short of 1) in Leitgeb’s truth condition, there will be a lottery with sufficiently many tickets to create a problem. The problem turns on the following ‘agglomeration’ principle, common to all of the leading logics of counterfactuals:

(Agglomeration) If \( p \rightarrow q \) and \( p \rightarrow r \), then \( p \rightarrow (q \& r) \).

Suppose, for example, that we set the bar at 0.999999. Consider a lottery with 1,000,001 tickets that in fact is never played. Then all of the following counterfactuals come out true, for their corresponding conditional probabilities clear the bar:

Lottery is played \( \rightarrow \) ticket 1 loses.

Lottery is played \( \rightarrow \) ticket 2 loses

…
Lottery is played □→ ticket 1,000,001 loses.

Then by Agglomeration, we have

Lottery is played □→ (ticket 1 loses & ticket 2 loses & … & ticket 1,000,001 loses).

This is equivalent to

Lottery is played □→ (each ticket loses).

But we also assume that

Lottery is played □→ some ticket wins.

By Agglomeration again, we conclude

Lottery is played □→ (each ticket loses and some ticket wins).

The antecedent is contingent, but the consequent is a contradiction, so this conclusion is a contradiction. The mistake, I claim, is to regard each of the ‘ticket i loses’ counterfactuals to be true. But by symmetry, they all have the same truth value. So they are all false, contra the Leitgeb truth conditions, with the bar for ‘very high’ set at 0.999999. And so it goes wherever we set the bar, short of 1. Moreover, most ordinary counterfactuals involve natural ‘lotteries’, as they are governed by chance processes. Even a dropped plate is not certain to fall; at best, it does so with very high chance.

Another way to respond is to deny Agglomeration. If Leitgeb sets the bar any lower than 1, this is what he must do. But that too is a significant cost, since Agglomeration is so intuitive. To uphold it, the bar must be set at 1—and by that, I do not mean “very high, or close to 1”. In fact, I think this even gives the intuitively correct verdicts for the individual lottery-ticket counterfactuals. Intuitively, it is false that ticket i would lose. What’s true is that ticket i would probably lose, but that is something else.
We thus come to a key move of Leitgeb’s, and we come full circle back to the difference between a proposition being true, and it being probably true, as I promised that we would. Recall Leitgeb’s claim of the near-synonymy of ‘would’ and ‘would very probably’, which I disputed. Previously, he regarded the truth condition for ‘if it were that $p$, it would be that $q$’ to be ‘$\text{ch}(q \mid p)$ is very high’, where this is understood to be above a vaguely determined threshold. In response to my lottery argument, he moves to regarding $\text{ch}(q \mid p)$ above an “exact threshold” to give the “semantics for approximate truth” (113). On this reading, each of the ‘individual ticket’ premises of the lottery argument is approximately true, but agglomeration fails, so the argument does not go through.

Likewise, I take it that on this reading, ‘if I had dropped the plate, it would have fallen to the floor’ is approximately true—though strictly speaking false. Leitgeb takes his cue from science, which is full of such statements. [PLEASE CORRECT ME IF I’M MISREADING HIM – I WAS TEARING MY HAIR OUT TRYING TO UNDERSTAND WHAT HE WAS UP TO HERE!]

I would be delighted if this worked: my uncompromising view about most counterfactuals would be vindicated, while commonsense would be mostly vindicated too—the truth of counterfactuals is exceedingly hard to come by, but approximate truth is rather easier. Indeed, for years I said much the same thing myself.

However, I eventually came to have misgivings with this way of speaking. In the end I preferred a cagier way of putting the point, which Leitgeb quotes (on p. 113): “In the neighborhood of the ordinary but false counterfactuals that we utter, there are closely related counterfactuals that are true but not ordinary. They are counterfactuals with appropriate probabilistic or vague consequents.” This is not to say that the false counterfactuals that we utter are approximately true, thanks to the truth of the related counterfactuals with appropriate probabilistic consequents. (Vague consequents are
A high chance of $p$ is not the same thing as $p$ being approximately true. Likewise, a high conditional chance of $q$, given $p$ is not the same thing as 'if it were that $p$, it would be that $q$' being approximately true.

For think of a case of approximate truth, taken from science. The speed of light is approximately 3 x $10^8$ m/s. Call its exact speed $c$. These two speeds are similar, and because of this, ‘the speed of light is 3 x $10^8$ m/s’ is approximately true. A world in which light has a speed of (exactly) 3 x $10^8$ m/s is ipso facto similar to our world in which it has speed $c$, with respect to light’s speed. But a world in which a coin lands heads with chance 0.99 is not ipso facto similar to a world in which it is true that the coin lands heads. The worlds need not be similar with respect to the coin's chance of heads: the latter world might have a low chance of heads, but it turns out to be true nonetheless. Nor need the worlds be similar with respect to the outcome: suppose the coin lands tails in the former world and heads in the latter. Similarly, a world in which the coin has a high conditional chance of landing heads, given it is tossed, is not ipso fact similar to a world in which it is true that the coin would land heads if it were tossed.

A way of thinking about this is that probability crosscuts similarity. As it turns out, Leitgeb himself comes close to making this point when discussing “the differences between Lewis’ [similarity-based] semantics and [his own] probabilistic semantics … after all, degrees of similarity are not supposed to correspond to magnitudes of probability in any straightforward manner” (55).

At least this allows me to end this section on a happy note of rapprochement. I have long argued against similarity accounts of counterfactuals, à la Stalnaker and Lewis. Worlds in which the plate is dropped and falls to the floor may well be more similar to ours than worlds in which it is dropped and does something else. But that doesn’t
make it true that if the plate were dropped, it would fall to the floor. (See my MS.) In
giving probabilistic semantics for counterfactuals, Edgington and Leitgeb must
likewise disavow similarity semantics. At least they are my allies in that regard.

4. Manifesto

It is election week in Australia as I write these words. There has been much sniping
among the politicians that those on the other side have not made clear what they stand
for. Let me say at least something about what I stand for with regard to counterfactuals.
Inspired by Karl Marx and Shane Warne, I want to finish with a brief manifesto of
some of my own views about counterfactuals.

A manifesto should begin with a slogan, and I will begin by repeating mine.

*Probability is probability of truth*. Truth is more fundamental than probability.
Counterfactuals make claims about how the world is; probabilities attach to such
claims. Counterfactuals represent the world as being certain ways rather than others;
probabilities attach to these representations being correct. So I am suspicious of no-
truth-value theories of counterfactuals, and doubly suspicious of *probabilistic* no-
truth-value theories of counterfactuals. I am sympathetic to a ‘probability first’
epistemology and theory of rationality, but unsympathetic to a ‘probability first’
metaphysics. More specifically, I have doubts about the prospects of such a
metaphysics of modality, and still more specifically, of such a metaphysics of
counterfactuals.

Counterfactuals play various roles in our mental lives that require them to have
truth values. It seems that they can be the objects of various factive attitudes:

He knows that if he were to strike the match, it would light.
She remembers that if she were to drink too much coffee, she would have trouble sleeping.

I reminded them that if they are late, they will miss out on the food…

And never mind factivity. It is hard to see how even non-factive attitudes can attach to things that have no truth value. You can’t believe a question, hope for an imperative, fear an optative, or desire an expletive. But you can bear all these attitudes to a counterfactual. Moreover, no-truth-value theories of conditionals generally, and of counterfactuals in particular, have trouble explaining the many ways in which counterfactuals combine compositionally with other parts of language, and how they iterate.

Counterfactuals can have various logical properties and bear various logical relations to other sentences. I have already mentioned incompatibility. They can be contradictory (recall the conclusion of the lottery argument in the last section), and they can be members of inconsistent sets of sentences (recall the partition of ‘Oswald’ counterfactuals in section 2.2 i)). They can appear as premises or as conclusions of valid arguments, not merely ‘probabilistically valid’ arguments. In fact, the validity of an argument cannot be recovered even as a limiting case of probabilistic validity. For \( P(B \mid A) = 1 \) is a weaker relation than \( A \) entailing \( B \), thanks to failures of regularity again. (Let \( A = \) the coin is tossed infinitely many times, and \( B = \) the coin does not land heads each time.) Edgington must deny that counterfactuals figure in such logical relations, while allowing that they figure in confirmation relations. This strikes me as just as unstable a combination as denying that they have truth values, while allowing that they have probabilities.

---

4 All right, depending on the expletive, *maybe* you can.
Moreover, if counterfactuals are to serve in accounts of causation, perception, knowledge, dispositions, rational decision, action, explanation, and so on, then counterfactuals had better have truth values. For statements about what causes what, who perceives what, who knows what, and so on, clearly do have truth values. It is mysterious how this could be if the counterfactuals that putatively ground them do not.

And most of these statements have objective truth values. (Statements about rational decision may be exceptional on this list, but their subjectivity arises not from counterfactuals, but from an agent’s attitudes to them.) The counterfactuals that subserve such statements must have objective truth values too. So I am suspicious of accounts of counterfactuals that ground them in subjective attitudes, such as credences, which renders mysterious how objectivity can emerge therefrom. And with two layers of probabilities—one subjective, one objective—I believe that Edgington’s account is twice removed from its target: the counterfactual itself.

Leitgeb strips away the subjective layer with his truth conditions (keeping the other one for his Skyrms-like assertability conditions). Objective-probabilistic accounts of counterfactuals, such as Leitgeb’s, are a step in the right direction: at least they accord counterfactuals objectivity. And according them truth values, as he does, is another such step. But I remain wary of inferences from probabilistic premises to probability-free conclusions—for I am no friend of regularity.

I am also wary of grounding counterfactuals in chance, for they are quite different kinds of modality. Chances are intimately tied to the laws of nature. Counterfactuals need not be; indeed, they need not even be tied to the laws of mathematics or logic. Counterfactuals are tied more closely to necessity and possibility—and these do not line up neatly with chances (witness irregularity yet again).
Yet there are, I believe, some valid inferences from claims about chances to claims about counterfactuals—but it’s the *falsehood* of those counterfactuals that are validly inferred. From $ch(q \mid p) > 0$, we may infer ‘if $p$ were the case, $q$ might be the case’; and from that, we may infer the falsehood of ‘if $p$ were the case, $q$ would not be the case’. This is rather like the inference from $ch(q) > 0$ to ‘$q$ is possible’, and from there to the falsehood of ‘not-$q$ is necessary’. In fact, I take the analogy between counterfactuals and necessity seriously. Indeed, I have a number of arguments that counterfactuals are strict conditionals. Little wonder that so many of them are false!

But that is for another occasion—see my MS. First the manifesto; then the revolution!