

# Counterfactual Entailment

David Barnett  
University of Colorado at Boulder

## ABSTRACT

*Counterfactual Entailment* is the view that a counterfactual conditional is true just in case its antecedent entails its consequent. I present an argument for *Counterfactual Entailment*, and I develop a strategy for explaining away apparent counterexamples to the view. The strategy appeals to the suppositional view of counterfactuals, on which a counterfactual is essentially a statement, made relative to the supposition of its antecedent, of its consequent.

Let us call a statement of the form ‘If  $A$  had been the case,  $C$  would have been the case’ a *counterfactual conditional*.<sup>1</sup> Under what conditions is a counterfactual conditional true?

Here is one idea:

*Counterfactual Entailment*      A counterfactual conditional is true if, and only if,  
its antecedent entails its consequent.

To ensure a unique reading of *Counterfactual Entailment*, let us agree that ‘entails’ means *metaphysically requires*. On this reading of ‘entails’, being made of water entails being made of H<sub>2</sub>O; having a hamster on your shoulder entails being such that two is an even number; and jumping out of a plane without a parachute does not entail getting injured (though it does make it likely).

*Counterfactual Entailment* gives rise to a puzzle.

On the one hand, it seems rife with counterexamples. Consider for instance the real life story of Philippe Petit. On the morning of August 7, 1974, Philippe and his friends secretly fixed a 200-kilogram cable between the roofs of the two 104-story World Trade Center buildings. Then, with no safety gear (other than an eight-meter balancing pole), Philippe walked back and forth eight times between the building tops. The question arises what would have happened if Philippe had fallen. Most of us are highly confident in the following answer:

(D)      If Philippe had fallen, he would have died.

Of course, it is metaphysically possible for Philippe to have fallen without dying. One way for Philippe to have fallen without dying is for a perfect gust of wind to have formed beneath him at just the right moment, bringing him to a gentle landing on his feet. To be sure, it is highly unlikely that such a gust would have formed, had Philippe fallen. But it *could* have formed; that is to say, it is metaphysically

---

<sup>1</sup> By this convention, a statement that, if  $A$  had been the case,  $C$  would have been the case, qualifies as counterfactual regardless of whether  $A$  is actually the case.

possible for it to have formed. Another way for Philippe to have fallen without dying is for there to have been a subtle difference in the initial conditions of the universe which led not only to Philippe's falling but also to the existence of a giant bed of feathers on the ground beneath him. Again, it is highly unlikely that such a scenario would have obtained, had Philippe fallen. But it could have. So, while the antecedent of (D) may make the consequent highly likely, it does not entail it. Yet we maintain a high degree of confidence in (D). Hence, (D) appears to be a counterexample to *Counterfactual Entailment*. Of course, (D) is not unique in this respect. On a daily basis we are willing to assert a wide range of counterfactuals whose antecedents obviously do not entail their consequents. We say such things as, 'If you had been wearing your seatbelt, you wouldn't have broken your nose,' and 'If we had taken the back road, we would have been on time,' and 'If I hadn't noticed that cyclist, you would have hit her.' Thus, *Counterfactual Entailment* seems rife with counterexamples.

On the other hand, *Counterfactual Entailment* appears to admit of a sound argument, which I present in §1. We are left with a puzzle.

To resolve the puzzle, one must explain away the apparent counterexamples or show that something is wrong with the argument. I do not know of any promising strategy for showing that something is wrong with the argument. I do however know of a promising strategy for explaining away the apparent counterexamples. The strategy appeals to the suppositional view of counterfactuals, on which a counterfactual is essentially a statement, made relative to the supposition of its antecedent, of its consequent. In §2, I present this strategy. In §3 I say why the apparent counterexamples cannot be explained away by appeal to a non-suppositional view. And in §4 I conclude that *Counterfactual Entailment* is a true principle whose apparent counterexamples should be explained away by appeal to the suppositional view.

### **1. An Argument for *Counterfactual Entailment***

Whether or not counterfactuals essentially involve suppositions, we may agree that there exists a practice of making and evaluating subjunctive claims in the context of—or relative to—subjunctive suppositions. For illustration, suppose that Philippe had fallen. In the context of this supposition, we can make various subjunctive claims, such as that Philippe would have died, that Philippe would not have been arrested, and that it would have been a sad day. We can also evaluate these subjunctive claims in various ways, for instance, as *probable*, *indeterminate*, or *true*. The question arises under what conditions it is objectively correct, in the context of a subjunctive supposition, to evaluate a subjunctive claim as *true*. The answer to this question provides the materials for an argument for *Counterfactual Entailment*.

In my experience, I have found that the easiest way for people to see the answer to this question is not directly, but rather through an examination of a different, though equivalent, question. The question concerns the conditions under which a *bet* on a subjunctive claim, made relative to a subjunctive supposition, qualifies as a *winning* bet. For illustration, imagine that, in the context of supposing that Philippe had fallen, you and I bet on whether Philippe would have died. I bet that he would not have died. You bet that he would have died. The loser must pay the winner 100 pounds. The question arises under what conditions your bet qualifies as a winning bet. Clearly, it is a winning bet just in case it is *true* that Philippe would have died. More generally, relative to a subjunctive supposition *that A had been the case*, the conditions under which a bet *that C would have been the case* qualifies as a winning bet are identical to the conditions under which the claim *that C would have been the case* qualifies as true. I have found that it is easiest for people to identify these conditions by first considering the question in terms of winning bets and then, only after answering the question in these terms, considering the question in terms of truth. For this reason, I formulate the crux of my argument in terms of winning bets rather than truth.<sup>2</sup> (Those who find talk of winning bets to be more of a distraction than an aid may reformulate the crux of my argument directly in terms of truth.)

---

<sup>2</sup> McDermott (1996) formulates his argument for a theory of *indicative* conditionals in terms of winning bets, rather than truth. I think this strategy is counterproductive when applied to indicatives; for an explanation, see my 2006.

We begin by considering the general question of what is necessary and sufficient for a bet that S to qualify as a winning bet. A simple answer is that a bet that S is a winning bet iff it is *true* that S. For present purposes, it will be useful to give a compatible answer that does not directly involve the notion of truth.

To see what I have in mind, imagine that Betty and Nancy have a dispute over whether any hamsters have blue eyes. Betty bets that some hamsters have blue eyes; Nancy bets that no hamsters have blue eyes. Under what conditions does Betty's bet qualify as a winning bet? Suppose that Betty discovers some facts about the typical genetic make-up of hamsters, and that these facts make it highly likely, but not certain, that some hamsters have blue eyes. Do these facts entail that Betty has won the bet? Clearly not. For Betty and Nancy have not bet on the question of whether it is highly likely that some hamsters have blue eyes; they have bet on the question of *whether some hamsters have blue eyes*. And the fact that it is highly likely that some hamsters have blue eyes does not settle this question. For Betty's bet to be a winning bet, it must be settled, not merely that it is *highly likely* that some hamsters have blue eyes, but *that some hamsters have blue eyes*. Obviously, the notion of *settling* that is relevant here is not epistemic. What is required for Betty's bet to be a winning bet is not for someone to *figure out* that some hamsters have blue eyes. Rather, it is for something to *make it the case* that some hamsters have blue eyes. In other words, something must *metaphysically guarantee* that some hamsters have blue eyes. Suppose, for instance, that there is a hamster whose eyes are in physical state P, and that having eyes in physical state P entails having blue eyes. Then something metaphysically guarantees that some hamsters have blue eyes, namely, the fact that some hamster has eyes in physical state P. Because this fact entails that some hamsters have blue eyes, and because Betty bet that some hamsters have blue eyes, Betty's bet is a winning bet. In general, then, for a bet that S to qualify as a winning bet, it is necessary and sufficient for it to be settled—that is, metaphysically guaranteed—that S.

Now consider an arbitrary subjunctive supposition:

(1) that A had been the case.

(1) is not the supposition that *A had been the case and everything else had been maximally similar to actuality*; it is simply the supposition that A had been the case. Imagine that, in the context of supposing (1), someone bets (2):

(2) that C would have been the case.

Under what conditions is (2) a winning bet? Clearly, it is not enough that (2) be *highly likely*, given (1). For emphasis, consider the following supposition:

Urn that yesterday you had randomly drawn from some urn or other containing ninety-nine red balls and one black ball.

In the context of supposing Urn, consider the following claims:

Red that you would have drawn red

Black that you would have drawn black.

You and I agree that Red is 99% likely, and that Black is 1% likely. Eager to bet on a long shot, I convince you to accept the following wager: you bet Red; I bet Black; the loser pays the winner 100 pounds. Now, does the mere fact that Red is highly likely entail that you have won the bet? Clearly not. For we have not bet on the question of whether it is highly likely that you would have drawn red; we have bet on the question of *what color you would have drawn*. And the fact that it is highly likely that you would have drawn red does not settle this question. In order for Red to be a winning bet, it must be

settled, not merely that it is highly likely that you would have drawn red, but *that you would have drawn red*.

Return now to our first bet. You bet that Philippe would have died; I bet that he would not have died. Does the mere fact that it is highly likely that Philippe would have died entail that you have won the bet? No. You and I might have agreed all along (i) that Philippe was not wearing any safety gear; (ii) that Philippe was a quarter mile above a paved sidewalk; and (iii) that, together with all the other facts, these facts make it *highly likely* that Philippe would have died. Indeed, it is precisely because I took it to be so *unlikely* that Philippe would have survived that I made the bet in the first place; I cannot resist a long shot. Obviously, then, the mere fact that it is highly likely that Philippe would have died does not entail that you have won the bet; for we have not bet on the question of whether it is highly likely that Philippe would have died. We have bet on *whether Philippe would have died*. And the fact that it is highly likely that Philippe would have died does not settle this question. In order for your bet to qualify as a winning bet, it must be settled, not merely that it is *highly likely* that Philippe would have died, but *that Philippe would have died*.

Again, the relevant notion of *settling* is not epistemic. What is required for your bet to be a winning bet is not for someone to *figure out* that Philippe would have died, but rather for something to *make it the case* that Philippe would have died. In other words, something must *metaphysically guarantee* that Philippe would have died.

If the supposition that Philippe had fallen *entails* that Philippe would have died, then something metaphysically guarantees that Philippe would have died, namely, *Philippe's having fallen*. The question arises whether something else—perhaps some contingent facts about Philippe's circumstance, together with the laws of nature—might metaphysically guarantee that Philippe would have died. To see that the answer is *no*, suppose (as you are already inclined to believe) that Philippe's having fallen does *not* metaphysically guarantee that he would have died. In other words, suppose that it is metaphysically possible for Philippe to have fallen without dying. Speaking again in the context of the supposition that Philippe had fallen, it follows that it is metaphysically possible for Philippe to have survived. Thus, it is not metaphysically guaranteed that Philippe would have died. Thus, nothing—including all the contingent facts about Philippe's circumstance, together with the laws of nature—metaphysically guarantees that Philippe would have died. We may conclude, then, that something metaphysically guarantees that Philippe would have died if, and only if, Philippe's having fallen metaphysically guarantees that Philippe would have died. In other words, something metaphysically guarantees that Philippe would have died if, and only if, the supposition that Philippe had fallen *entails* that Philippe would have died. Because your bet that Philippe would have died is a winning bet only if something metaphysically guarantees that Philippe would have died, your bet is a winning bet only if the supposition that Philippe had fallen entails that Philippe would have died.

The preceding reasoning constitutes the key move in my defense of *Counterfactual Entailment*. Because the move is invalid in the case of indicatives, some readers might mistakenly deem it invalid in the present case. Indeed, a common reaction to the move is to reconstruct it in abstraction from the subjunctive character of the supposition *that Philippe had fallen* and the claim *that Philippe would have died*, by substituting sentence letters for the two that-clauses; to then deem the reconstruction invalid; and to then deem the original move invalid on that basis. Because the validity of the move depends essentially on the subjunctive character of the supposition and the claim, such reconstructions will indeed appear invalid. It is hard for me to emphasize enough that one cannot properly evaluate the key move by considering analogues in which the subjunctive character has been lost.

To help avoid this confusion, I now apply the move, step by step, to show that (2) is a winning bet only if (1) entails (2). At each step, the reader should pay careful attention to the subjunctive natures of (1) and (2). After I apply the move here, I will address a tempting objection to the move. Then I will show how the move fails in the case of indicatives. The move can be broken into seven steps:

- Step 1: Suppose that (1) does not entail (2). In other words, suppose that *A*'s having been the case does not metaphysically guarantee that *C* would have been the case.

- Step 2: Then it is metaphysically possible for A to have been the case without C's having been the case.
- Step 3: So, on the supposition that A had been the case, it is metaphysically possible that C would not have been the case.
- Step 4: So, on the supposition that A had been the case, it is not metaphysically guaranteed that C would have been the case.
- Step 5: So, on the supposition that A had been the case, *nothing* metaphysically guarantees that C would have been the case.
- Step 6: Hence, something metaphysically guarantees that C would have been the case *only* if (1) entails (2).
- Step 7: Because (2) is a winning bet only if something metaphysically guarantees that C would have been the case, (2) is a winning bet only if (1) entails (2).

Steps 1 – 7 constitute the key move in my defense of *Counterfactual Entailment*. Here is a tempting objection to the move:

Let 'B' abbreviate the claim that the sky is blue. It is metaphysically possible that not-B. Applying the move from Step 3 to Step 4, we infer that it is not metaphysically guaranteed that B. Applying the move from Step 4 to Step 5, we infer that nothing metaphysically guarantees that B. Of course, B is true only if something metaphysically guarantees that B. Hence, B is not true. So, from the truth that it is metaphysically possible that the sky is not blue, we have inferred the falsity that it is not true that the sky is blue, by applying the reasoning from Step 3 to Step 5. Hence, this reasoning is invalid.<sup>3</sup>

The appeal of this objection rests on a failure to distinguish (3a) from (3b):

- (3a) It is metaphysically possible that the sky is not blue.
- (3b) It is metaphysically possible for the sky to have not been blue.

Philosophers are trained to miss the distinction between (3a) and (3b). As students, we are told that 'It is metaphysically possible that a is F' is true just in case it is metaphysically possible for a to have been F, whether or not a is in fact F. While talking this way may serve various pedagogical and even certain philosophical purposes, strictly speaking, it is incorrect. For some people, it might be *epistemically* possible that the sky is—in fact—not blue. But, as a matter of fact, it is *not* metaphysically possible that the sky is—in fact—not blue. (I say 'in fact' only to emphasize the indicative nature of the claim.) Do not confuse this claim with the claim that it is not metaphysically possible for the sky to *have not been* blue. The first claim is equivalent to the claim that things are such that it is metaphysically impossible that the sky is—in fact—not blue. And things *are* that way; for, as a matter of fact, the sky *is* blue; and, if the sky is blue, then it is not metaphysically possible that the sky is not blue. The second claim is equivalent to the claim that things are such that it is metaphysically impossible for the sky *to have not been* blue. And

<sup>3</sup> I am indebted to Michael Huemer for raising this objection.

things are *not* that way; for, as a matter of fact, the sky *could have not been* blue. (Hereafter, I use ‘could’ to designate metaphysical possibility.) Despite what we are told as students, the claim that it is metaphysically impossible that the sky *is* not blue does not entail the claim that it is metaphysically impossible for the sky *to have not been* blue.

An analogy might be useful. Suppose that you do not currently have a driver’s license. Then something excludes the legal possibility of your driving. Does it follow that it is legally impossible for you to drive, right now? Perhaps there is an initial temptation to say *no*. But this is not right. To make it possible for you to drive legally, something must change. As things currently stand, it is *not* legally possible for you to drive. As things currently stand, it is legally impossible for you to drive. Do not confuse this truth with any of the following falsehoods:

- (4a) As things currently stand, it is legally impossible for you to drive *in the future*.
- (4b) As things currently stand, it is legally impossible for you to *have driven in the past*.
- (4c) As things currently stand, it is legally impossible for you to *have driven right now, had things been different*.

Supposing, then, that you do not currently have a driver’s license, it is legally impossible for you to drive *right now*.

Returning to (3a) and (3b), suppose that the sky is blue. Then something excludes the metaphysical possibility that the sky is not blue. Does it follow that it is metaphysically impossible that the sky is not blue? Perhaps there is an initial temptation to say *no*. But this is not right. As things currently stand, it is *not* metaphysically possible that the sky is—at this moment—not blue. Do not confuse this truth with any of the following falsehoods:

- (3c) As things currently stand, it is metaphysically impossible that the sky *will not be blue in the future*.
- (3d) As things currently stand, it is metaphysically impossible that the sky *was not blue in the past*.
- (3e) As things currently stand, it is metaphysically impossible that the sky *would not, at this very moment, have been blue, had things been different*.

Supposing, then, that the sky is blue, it is metaphysically impossible that the sky is not blue.

Some things, such as the color of the sky and your legal status as a driver, could have been different from the way they are. In other words, it is metaphysically possible for some things to have been different from the way they are. It is not metaphysically possible, however, that anything is different from the way it is. If something is a certain way, then while it may be epistemically possible that it is not that way, it is not metaphysically possible that it is not that way. In general, if *a* is *F*, then it is metaphysically impossible that *a* is not *F*. To be sure, it may nevertheless be metaphysically possible (i) that *a* will not be *F* in the future; (ii) that *a* was not *F* in the past; and (iii) that *a* would not currently have been *F*, had things been different. But it is not metaphysically possible that *a is not F*.<sup>4</sup>

---

<sup>4</sup> Given that *a* is *F*, is it metaphysically possible for *a* not to be *F*? This question is formulated in the present subjunctive—to be *F*—which, despite its name, is tenseless. Unless the question involves an implicit reference to the present tense—that is, unless the real question is whether it is metaphysically possible for *a* not to be *F* *right now*—the answer to the question might be *yes*. For, given that *a* is *F*, it might be metaphysically possible for *a* not to be *F* *in the future*.

At this point in the dialectic, I have heard it suggested that all I have shown is that there is a fact—namely, that the sky is blue—that is metaphysically incompatible with the truth of the claim that the sky is not blue. This, it is suggested, does not establish the substantive claim (a) that it is metaphysically impossible that the sky is not blue. Rather, it merely establishes the trivial claim (b) that it is metaphysically impossible that the sky is not blue, *given that the sky is blue*. But this suggestion is misguided. For the sky *is* blue. Given (b), (a) follows.

The objection that we have been considering to the key move rests on the assumption that (3a)—it is metaphysically possible that the sky is not blue—is true. As philosophers, we are trained to find this assumption appealing, for we are trained to miss the distinction between (3a) and (3b). On reflection, however, we see that, whereas (3b) is true, (3a) is false. Hence, the objection fails.

To emphasize that the key move depends essentially on the subjunctive natures of (1) and (2), I will now show where the move goes wrong when applied to the case of a bet on an indicative claim, made in the context of an indicative supposition. In other words, I will show why the move cannot be used to establish that such a bet is a winning bet only if it is entailed by the corresponding supposition.

Consider an arbitrary indicative supposition:

(1\*) that A is the case.

In the context of this supposition, someone makes the following bet:

(2\*) that C is the case.

Now we attempt to apply the key move:

- Step 1: Suppose that (1\*) does not entail (2\*). In other words, suppose that A's being the case does not metaphysically guarantee that C is the case.
- Step 2: Then it is metaphysically possible for A to be the case without C's being the case.
- Step 3: So, on the supposition that A is the case, it is metaphysically possible that C is not the case.
- Step 4: So, on the supposition that A is the case, it is not metaphysically guaranteed that C is the case.
- Step 5: So, on the supposition that A is the case, *nothing* metaphysically guarantees that C is the case.
- Step 6: Hence, something metaphysically guarantees that C is the case *only* if (1\*) entails (2\*).
- Step 7: Because (2\*) is a winning bet only if something metaphysically guarantees that C is the case, (2\*) is a winning bet only if (1\*) entails (2\*).

To see where the reasoning goes wrong, suppose that some hamsters do in fact have blue eyes. Clearly, this does not entail that some gerbils have blue eyes. May we infer that it is metaphysically possible that no gerbils have blue eyes? No. For, consistent with our supposition, it might be the case that some gerbils do in fact have blue eyes. And if it is the case that some gerbils do in fact have blue eyes, then it is not metaphysically possible that no gerbils in fact have blue eyes. (To be sure, it might be metaphysically possible (a) that, in the past, no gerbils had blue eyes; (b) that, in the future, no gerbils will have blue eyes; and (c) that no gerbils would, presently, have had blue eyes, had things been

different.) So, from the fact that our supposition does not entail that some gerbils have blue eyes, we are not justified in inferring that it is metaphysically possible that no gerbils have blue eyes. All that we may infer is that *our supposition* does not exclude the metaphysical possibility that no gerbils have blue eyes. It is an open question whether something *else*—such as the fact that some gerbils have blue eyes—excludes this possibility.

So, applied to the case of (1\*) and (2\*), the key move goes wrong from Step 2 to Step 3. Unlike the subjunctive case of (1) and (2), in the indicative case something *other* than the supposition might exclude the bet from being a metaphysical possibility—namely, the actual facts. Hence, the key move does indeed depend for its validity on the subjunctive natures of (1) and (2).

Let us return to my main line of argument. So far, I have argued that, in the context of a supposition *that A had been the case*, a bet *that C would have been the case* is a winning bet only if *A*'s having been the case *entails* that *C* would have been the case. Of course, if *A*'s having been the case entails that *C* would have been the case, then such a bet is a winning bet. So such a bet is winning bet if, and only if, *A*'s having been the case entails that *C* would have been the case.

Now, a bet that S is a winning bet if, and only if, it is *true* that S. Returning to our original question, we may conclude that, in the context of a supposition *that A had been the case*, a claim *that C would have been the case* is *true* if, and only if, *A*'s having been the case *entails* that *C* would have been the case.

Now consider (5a) – (5c):

(5a) Suppose that Philippe had fallen. He would have died.

(5b) Supposing that Philippe had fallen, he would have died.

(5c) If Philippe had fallen, he would have died.

It would be bizarre for someone to deem the second sentence in (5a) true and either (5b) or (5c) not true. Likewise, it would be bizarre for someone to deem (5b) or (5c) true and the second sentence in (5a) not true. More generally, it would be bizarre to assign one truth-value to the second sentence of (6a) and a different truth-value to (6b) or (6c):

(6a) Suppose that *A* had been the case. *C* would have been the case.

(6b) Supposing that *A* had been the case, *C* would have been the case.

(6c) If *A* had been the case, *C* would have been the case.

Thus, it should be uncontroversial that the second sentence of (6a) always has the same truth-value as the corresponding sentences, (6b) and (6c). Applying our truth-conditions for the second sentence of (6a) to (6c), we may conclude that a statement *that, if A had been the case, C would have been the case* is true if, and only if, the antecedent entails the consequent. In other words: *Counterfactual Entailment*.

My argument for *Counterfactual Entailment* consists, then, of the key move together with several additional steps which jointly identify (i) the conditions under which a subjunctive bet, made relative to a subjunctive supposition, qualifies as a *winning* bet, with (ii) the conditions under which a corresponding counterfactual conditional statement qualifies as *true*. Here is the complete argument:

Step 1: Suppose that (1) does not entail (2). In other words, suppose that *A*'s having been the case does not metaphysically guarantee that *C* would have been the case.

Step 2: Then it is metaphysically possible for *A* to have been the case without *C*'s having been the case.

- Step 3: So, on the supposition that A had been the case, it is metaphysically possible that C would not have been the case.
- Step 4: So, on the supposition that A had been the case, it is not metaphysically guaranteed that C would have been the case.
- Step 5: So, on the supposition that A had been the case, *nothing* metaphysically guarantees that C would have been the case.
- Step 6: Hence, something metaphysically guarantees that C would have been the case *only* if (1) entails (2).
- Step 7: Because (2) is a winning bet only if something metaphysically guarantees that C would have been the case, (2) is a winning bet only if (1) entails (2).
- Step 8: If (1) entails (2), then (2) is a winning bet.
- Step 9: So, (2) is a *winning* bet iff (1) entails (2).
- Step 10: Because (2) is a winning bet iff (2) is true, (2) is *true* iff (1) entails (2).
- Step 11: (2) is true iff the following statement is true: if A had been the case, C would have been the case.
- Step 12: Hence, *Counterfactual Entailment*.

We are left with a puzzle. On the one hand, *Counterfactual Entailment* seems rife with counterexamples. On the other, *Counterfactual Entailment* seems to admit of a sound argument. To resolve the puzzle, one must explain away the apparent counterexamples or show that something is wrong with the argument. As I said in my opening remarks, I know of no promising strategy for showing that something is wrong with the argument, but I do know of a promising strategy for explaining away the apparent counterexamples.

## 2. A strategy for explaining away apparent counterexamples to *Counterfactual Entailment*

The strategy has four components.

The first component is the suppositional view of counterfactuals. On this view, to state *that, if A had been the case, C would have been the case* is, essentially, to state something relative to a supposition: what is stated is *that C would have been the case*; what is supposed is *that A had been the case*. To see that this is the natural view, consider (5b) and (5c) again:

(5b) Supposing that Philippe had fallen, he would have died.

(5c) If Philippe had fallen, he would have died.

To the untutored ear, (5b) and (5c) sound synonymous. The On-line Oxford English Dictionary agrees that they are synonymous; it defines ‘if’ as follows: ‘On the condition that; given or granted that; in (the) case that; supposing that; on the supposition that.’ It is no coincidence that if-then statements are called *conditional* statements: they appear to be statements of their consequents *conditional on*—that is, *relative to the supposition of*—their antecedents. It is natural, then, to treat ‘if A had been the case’ as synonymous with ‘supposing that A had been the case’—both clauses function to express the supposition that A had

been the case. Likewise, it is natural to treat (5b) and (5c) as synonymous: both state *that Philippe would have died* relative to the supposition *that Philippe had fallen*.

On the rival *categorical* view, to state *that, if A had been the case, C would have been the case* is to state something, not relative to any hypothesis or supposition, but *categorically*. Proponents of the categorical view disagree over *which* thing is categorically stated by a counterfactual. Nelson Goodman (1947) says that it is an entailment from the antecedent, together with laws of nature and particular facts about the actual world, to the consequent;<sup>5</sup> Robert Stalnaker (1968) says that it is a predication of a single possible world; and David Lewis (1973) says that it is an existential generalization over a set of possible worlds. By contrast, W.V.O. Quine (1950), John Mackie (1973), Michael Dummett (1978), Dorothy Edgington (1995), and I (2006, 2009, 2010) maintain that counterfactuals are as their surface form suggests: *suppositional* statements.<sup>6</sup> On this view, nothing is categorically stated by a counterfactual. Counterfactual statements are essentially acts of supposing-cum-stating.<sup>7</sup>

The second component of the strategy is the principle that, to assign a value to a counterfactual statement is to assign the value to what is stated by the consequent of the counterfactual, relative to what is supposed by the antecedent. For instance, to assign a probability to a counterfactual is to assign the probability to what is stated by the consequent, relative to what is supposed by the antecedent. So, to say that it is 99% probable *that, if Philippe had fallen, he would have died* is to assign 99% probability to the claim *that Philippe would have died*, relative to the supposition *that Philippe had fallen*. Likewise, to have a degree of confidence in a counterfactual is to have a degree of confidence in what is stated by the consequent, relative to what is supposed by the antecedent. And to ascribe the value *true* to a counterfactual is to ascribe the value *true* to what is stated by the consequent, relative to what is supposed by the antecedent.

The third component is the principle that, relative to a subjunctive supposition, an ascription of a value to a subjunctive claim commits one, not to the proposition that the claim *has* the value, but rather to the proposition that the claim *stands in a corresponding relation* to the supposition. For instance, in the context of the supposition *that Philippe had fallen*, an ascription of 99% probability to the claim *that Philippe would have died* commits one, not to the proposition that what is claimed has an absolute probability of 99%, but rather to the proposition that what is claimed *is made 99% probable* by what is supposed. And, in the context of the supposition *that Philippe had fallen*, an ascription of the value *true* to the claim *that Philippe would have died* commits one, not to the proposition that what is claimed is absolutely true, but rather to the proposition that what is claimed is *made true*, and thus *entailed*, by what is supposed.

The fourth component centers on the fact that the antecedent of a counterfactual can make the consequent highly probable without entailing it. Given this fact, together with the preceding principles, it is easy to see that a high degree of confidence in a counterfactual is rationally consistent with zero degree of confidence that the counterfactual is true. A high degree of confidence in a counterfactual commits one to the proposition that the antecedent makes the consequent highly probable; zero degree of confidence that a counterfactual is true commits one to the proposition that the antecedent does not entail the consequent; and there is no problem in the idea of a counterfactual whose antecedent makes its consequent highly probable without entailing it.<sup>8</sup> Applied to (D), one can rationally maintain a high degree of confidence in (D) together with zero degree of confidence that (D) is true. One might express this combination of mental states as follows: ‘I’m highly confident that Philippe would have died, if he had fallen. I’m also certain that there is no objective fact of the matter whether he would have died. For

---

<sup>5</sup> More recent views inspired by Goodman’s account include Pollock 1981, Barker 1999, and Hiddleston 2005.

<sup>6</sup> Other views consistent with the suggestion include Adams 1965, 1966, 1975, Ryle 1950, and Woods 1997.

<sup>7</sup> I have heard it suggested that the categorical view is compatible with the suppositional view. The suggestion is false, for the two views make a host of incompatible predictions concerning the correct evaluation of counterfactuals. For discussion of some incompatibilities in prediction, see my 2006 and 2010.

<sup>8</sup> I should add that there are no further jointly problematic commitments of a high degree of confidence in a counterfactual and a low degree of confidence that the counterfactual is true.

although the facts make it *highly likely* that he would have died, and *highly unlikely* that he would have survived, they do not *guarantee* that he would have died, nor do they *guarantee* that he would have survived. And so, strictly speaking, the facts neither verify nor falsify the claim that he would have died.<sup>7</sup>

We are now positioned to explain away the apparent counterexamples to *Counterfactual Entailment*. The appearance of counterexamples derives from a tempting, though invalid, transfer of confidence from one sort of claim to another. We are justifiably confident in various counterfactual statements whose antecedents obviously do not entail their consequents. Then, in each case, we mistakenly transfer this confidence to the corresponding claim, of the given counterfactual statement, that it qualifies as *true*, and thus that it qualifies as a counterexample to *Counterfactual Entailment*. Our disposition to transfer our confidence from a counterfactual statement to an ascription of truth to that statement perhaps derives from our disposition to make the *valid* transfer from categorical statements to ascriptions of truth to those statements. For instance, if one is justified in being highly confident that some hamsters bite, then one is justified in being highly confident that it is true that some hamsters bite. Paradoxes aside, the transfer of confidence from categorical statements to ascriptions of truth to those statements is valid. Perhaps we are conditioned by operating in categorical settings to *always* transfer confidence from a statement to an ascription of truth to that statement. This conditioning would explain why we are tempted to transfer our confidence in counterfactual statements whose antecedents obviously do not entail their consequents to ascriptions of truth to those statements.<sup>9</sup>

Applied to (D), this explanation predicts a temptation to transfer our justifiably high degree of confidence in (D) to an ascription of truth to (D), and thus to the claim that (D) is a counterexample to *Counterfactual Entailment*. But it would be a mistake to transfer our confidence in this way. For, whereas a high degree of confidence in (D) merely commits us to the claim that the antecedent of (D) makes the consequent *highly likely*, a high degree of confidence in an ascription of truth to (D) commits us to the claim that it is highly likely that the antecedent of (D) *entails* the consequent. And the antecedent of (D) obviously does not entail the consequent. So, we are positioned to see both (i) why we are tempted to transfer our justifiably high degree of confidence in (D) to the claims that (D) is true and that (D) is a counterexample to *Counterfactual Entailment*; and (ii) why this temptation should be resisted.<sup>10</sup>

So, we have available a promising strategy for explaining away the apparent counterexamples to *Counterfactual Entailment*. The strategy centers on the suppositional view of counterfactuals. Given this view, it is plausible that the appearance of counterexamples to *Counterfactual Entailment* derives from a

---

<sup>9</sup> Personally, I have never been tempted to transfer my confidence in this way. I still recall my first exposure to the philosophical debate over counterfactuals. I was taking an undergraduate course in metaphysics. Professor Luc Bovens: ‘Class, what is your reaction to the following statement: If Nixon had pressed the button, there would have been a nuclear holocaust?’ Me: ‘Probably.’ Professor Bovens: ‘You know the rules, Barnett. Answer *true* or *false*. Which is it?’ Me: ‘I don’t feel comfortable saying *true* or saying *false*. I feel comfortable saying *probably*.’ At the time, I was puzzled over why I did not feel comfortable saying *true*. After all, I recognized that I was highly confident that, if Nixon had pressed the button, there would have been a nuclear holocaust. And I recognized that, in other matters, certainty was not a prerequisite for my labeling something ‘true’; for instance, I was less than certain that I was not dreaming, but this did not prevent me from saying that it was true that I was not dreaming. So what was preventing me from saying that it was true that, if Nixon had pressed the button, there would have been a nuclear holocaust?

<sup>10</sup> What about the fact that we are willing on a daily basis to *assert* counterfactuals whose antecedents obviously do not entail their consequents? This fact should not compel us to conclude that such counterfactuals constitute counterexamples to *Counterfactual Entailment* either; for this willingness may simply reflect our high degrees of confidence in the counterfactuals. And it would be a mistake to transfer these high degrees of confidence to ascriptions of truth to the counterfactuals.

Williamson (1996) argues, roughly, that for a sincere and literal statement to be appropriate, its author must take herself to *know* what is stated; mere confidence in what is stated is not enough. This might be right for categorical statements, but not for subjunctive suppositional statements. For a discussion of the norm of assertion for counterfactuals, on the suppositional view, see my forthcoming (b).

tempting, though invalid, transfer of confidence from counterfactuals to ascriptions of truth to those counterfactuals.<sup>11</sup>

### 3. Other strategies for explaining away apparent counterexamples to *Counterfactual Entailment*

Might there be promising alternative strategies for explaining away the appearance of counterexamples to *Counterfactual Entailment*? Perhaps there are promising alternative *suppositional* strategies, but it is hard to see how there could be any promising alternative *categorical* strategies.

In the first place, it is hard to see how any plausible categorical view could be compatible with my argument for *Counterfactual Entailment*. To begin to see the difficulty, suppose that a counterfactual is a categorical statement. The question arises which thing is categorically stated by a counterfactual. Given *Counterfactual Entailment*, one might think that what is categorically stated is:

*Entailment*      that *A*'s having been the case *entails* that *C* would have been the case.

But, given my argument for *Counterfactual Entailment*, this cannot be right. For the core reasoning of the argument can be used, as follows, to establish that the *falsity* conditions of counterfactuals are distinct from the falsity conditions of *Entailment*.

Consider an arbitrary subjunctive supposition:

(1)      that *A* had been the case.

Imagine that, in the context of supposing (1), someone bets (2):

(2)      that *C* would have been the case.

Under what conditions is (2) a *losing* bet? For (2) to be a losing bet, it is not enough that something make it *highly unlikely* that *C* would have been the case. Rather, something must *metaphysically guarantee* that *C* would *not* have been the case. For something to *metaphysically guarantee* that *C* would not have been the case is for something to *exclude the metaphysical possibility* that *C* would have been the case. But for this possibility to be excluded, on the supposition that *A* had been the case, is for *A*'s having been the case to be metaphysically *incompatible* with *C*'s having been the case. Hence, for (2) to be a losing bet, on the supposition that *A* had been the case, is for this supposition to be metaphysically incompatible with *C*'s having been the case. Of course, given (1), (2) is a losing bet iff (2) qualifies as false. So, given (1), (2) qualifies false iff (1) is metaphysically incompatible with (2). By employing essentially the same final reasoning of our original argument, we may infer:

*Counterfactual Incompatibility*      A counterfactual conditional is *false* if, and only if, its antecedent is metaphysically incompatible with its consequent.

---

<sup>11</sup> A proponent of the suppositional view must address two questions. First, how should we treat compound statements with conditional statements as parts? Second, how should we characterize valid reasoning with conditionals? For indicative conditionals, I treat these questions in some detail in my 2006. (For similar treatments, see Adams 1965, 1975, 1998, and Edgington 1995; for discussion, see Bennett 2003.) Because my strategy for treating subjunctives is generally the same, for the sake of brevity, I refer the reader to my 2006. For present purposes, what is most important to note about the logic of conditionals, on the suppositional view, is that the key notion of validity is given, not in terms of truth-preservation, but rather in terms of Adams's notion of *probabilistic validity*, where an argument is *probabilistically valid* iff there is no probability function in which the uncertainty of the conclusion exceeds the sum of the uncertainties of the premises. The notion of probabilistic validity provides the basis for a formal logic of conditionals and other suppositional statements. A noteworthy feature of the logic is that, for arguments not involving suppositional statements, an argument is probabilistically valid iff it is truth-preserving; and so classical truth-functional logic is preserved.

Given *Counterfactual Entailment* and *Counterfactual Incompatibility*, it cannot be generally correct that a counterfactual is a categorical statement of *Entailment*. For a categorical statement is true/false iff what it categorically states is true/false. And, whereas *Entailment* is false iff *A*'s having been the case *does not entail* that *C* would have been the case, a counterfactual is false iff *A*'s having been the case is *metaphysically incompatible with C*'s having been the case. Because it is possible for *A*'s having been the case to be metaphysically compatible with, but not entail, *C*'s having been the case, it cannot generally be correct that a counterfactual is a categorical statement of *Entailment*.

Given *Counterfactual Entailment* and *Counterfactual Incompatibility*, it is hard to see what could be stated categorically by a counterfactual. For, by these two principles, there is no single condition such that a counterfactual is *true* under that condition and *false* otherwise. It is tempting to infer that a counterfactual cannot be a categorical statement of anything; for if it were, there *would* be a single condition of the preceding sort, namely, the truth-condition of the thing that was categorically stated by the counterfactual. But perhaps this inference is insensitive to the possibility of indeterminacy in the meanings of counterfactuals. Perhaps it is highly indeterminate *which* thing is categorically stated by a counterfactual, and perhaps this indeterminacy breeds just the sort of truth-value gap jointly entailed by *Counterfactual Entailment* and *Counterfactual Incompatibility*.

Applied to (D), the proposal might go as follows. First, there is a wide range of candidate propositions, each such that it is indeterminate whether *it* is what is categorically stated by (D). Second, all of these candidates are true on the condition that the antecedent of (D) entails the consequent, and all are false on the condition that antecedent of (D) is incompatible with the consequent. Third, on the condition that the consequent of (D) is neither entailed by, nor incompatible with, the antecedent, some of the candidates are true and some are false. Fourth, applying the method of supervaluations, a counterfactual statement is *true* on the condition that the antecedent of (D) entails the consequent; *false* on the condition that antecedent of (D) is incompatible with the consequent; and *neither true nor false* otherwise. In this way, the categorical view can be made to comport with both *Counterfactual Entailment* and *Counterfactual Incompatibility*.

The proposal is, however, counterintuitive. To see that it is counterintuitive, we need to set aside irrelevant sources of indeterminacy and consider whether, intuitively, there exists the sort of indeterminacy posited by the proposal. To do this, let us imagine away three irrelevant candidate sources of indeterminacy in (D). First, of all the perfectly precise sets of fundamental particles that are plausible candidates for constituting Philippe's body (as he crosses the tightrope), let us imagine that one is such that 'Philippe' determinately refers to it. Second, of all the perfectly precise sets of conditions that are plausible candidates for constituting the application conditions for 'falls', let us imagine that one is such that the application conditions of 'falls' are determinately given by it. Third, let us imagine the same for 'dies'. We shall thereby imagine away any possible indeterminacy in the meanings of 'Philippe', 'falls', and 'dies'.

In this context, suppose that Philippe had fallen. Would he have died? Here it seems that there is no indeterminacy as to what question is being asked. To be sure, the answer to this question is not settled by the supposition. And so, in the context of the supposition, we may label the question *indeterminate*. However, in doing this, we do not commit to the claim that the question is *absolutely* indeterminate—that is, absolutely unsettled—but rather to the claim that the question is *unsettled by what is supposed*. This is consistent with the intuition that there is no indeterminacy as to *which* question is being asked.

Now, while still imagining away any possible indeterminacy in the meanings of 'Philippe', 'falls', and 'dies', consider the corresponding counterfactual conditional: Would Philippe have died if he had fallen? Again, there does not seem to be any indeterminacy as to *which* question is being asked here. To be sure, there may be a temptation to ascribe indeterminacy to the counterfactual. But what is the source of this temptation? Is it that there appears to be indeterminacy as to *which* question is under consideration in the first place? Or is it rather that there appears to be a perfectly determinate question before us, but that its answer is not settled by the supposition relative to which it is asked? Intuitively, it is the latter, rather than the former, which is the source of the temptation. This is a problem for the

categorical proposal under consideration. For by it, even after we imagine away any indeterminacy in the meanings of ‘Philippe’, ‘falls’, and ‘dies’, there is still indeterminacy as to *which* question is the question of whether, if Philippe had fallen, he would have died. And this seems wrong.

The main question that we are considering in this section is whether the apparent counterexamples to *Counterfactual Entailment* might be explained away by appeal to a categorical view. So far I have argued that no plausible version of the categorical view is even compatible with my argument for *Counterfactual Entailment*; for, the central reasoning of this argument can be used to establish *Counterfactual Incompatibility*; and no plausible version of the categorical view is compatible with both *Counterfactual Entailment* and *Counterfactual Incompatibility*.

By contrast, the suppositional strategy clearly is compatible with both *Counterfactual Entailment* and *Counterfactual Incompatibility*. By the first two components of this strategy, an ascription of truth/falsity to a counterfactual is an ascription of truth/falsity to what is stated by the counterfactual, relative to what is supposed. By the third component, an ascription of truth/falsity to what is stated by the counterfactual, relative to what is supposed, commits one to the claim that what is stated is *made true/false* by what is supposed, and is thus *entailed by/incompatible with* what is supposed. Thus, on the suppositional strategy, it is objectively correct to ascribe *true/false* to a counterfactual iff what is stated is *entailed by/incompatible with* what is supposed. In other words, the suppositional strategy predicts both *Counterfactual Entailment* and *Counterfactual Incompatibility*.

Here is a second reason to doubt that the apparent counterexamples to *Counterfactual Entailment* might be explained away by appeal to a categorical view. Once *Counterfactual Entailment* is granted, the categorical view of counterfactuals demands that we have *zero* degree of confidence in every counterfactual for which we have zero degree of confidence that its antecedent entails its consequent. This is because, on the categorical view, a degree of confidence in a counterfactual is a degree of *absolute* confidence in what is categorically stated by the counterfactual, and a degree of confidence in an ascription of truth to a counterfactual is a degree of absolute confidence in the truth of what is categorically stated by the counterfactual. A degree of absolute confidence in the truth of *p* rationally requires an equal degree of absolute confidence in *p*. In particular, zero degree of absolute confidence in the truth of *p* rationally requires zero degree of absolute confidence in *p*. Given *Counterfactual Entailment*, certainty that *the antecedent of a given counterfactual does not entail the consequent* rationally requires zero degree of confidence in the truth of the counterfactual. Given the categorical view, this zero degree of confidence in turn rationally requires zero degree of confidence in the counterfactual.

But certainty that *Philippe’s having fallen does not entail that he would have died* clearly is rationally compatible with a high degree of confidence that, if Philippe had fallen, he would have died. Likewise, certainty that *your having randomly drawn from an urn containing ninety-nine red balls and one black ball does not entail that you would have drawn red* clearly is rationally compatible with a high degree of confidence that, if you had so drawn, you would have drawn red. So, given *Counterfactual Entailment*, the categorical view of counterfactuals is implausible. This is a second reason to doubt that the apparent counterexamples to *Counterfactual Entailment* might be explained away by appeal to a categorical view.

One might respond by objecting in similar fashion to the idea that *Counterfactual Entailment* might admit of a promising argument in the first place:

Certainty that *Philippe’s having fallen does not entail that he would have died* clearly is rationally compatible with a high degree of confidence that it is *true* that, if Philippe had fallen, he would have died. So *Counterfactual Entailment* is implausible. Hence, it is hard to see how there could be any promising argument for *Counterfactual Entailment*.

But this objection is not dialectically symmetrical to the one I gave in the preceding paragraph. For we already have before us a promising argument for *Counterfactual Entailment*, together with a promising strategy for explaining away just the sort of appearances on which the current objection to the possibility of such an argument is based. By contrast, we do not have before us a promising categorical strategy for explaining away the apparent counterexamples to *Counterfactual Entailment*, nor do we have before us a promising strategy for explaining away the appeal of my objection to the possibility of such a strategy. And so the two objections are not on a par: my objection has significant dialectical force; the current objection does not.

I conclude that the apparent counterexamples to *Counterfactual Entailment* cannot be explained away by appeal to a categorical view. My argument for *Counterfactual Entailment* thus stands as an argument for the suppositional view of counterfactuals, which provides for the only plausible strategy for explaining away the apparent counterexamples to *Counterfactual Entailment*.<sup>12</sup>

#### 4. Conclusion

We began with a puzzle about *Counterfactual Entailment*. On the one hand, it appeared to be rife with counterexamples. On the other, it appeared to admit of a sound argument (§1). To resolve the puzzle, I appealed to the suppositional view of counterfactuals to explain away the apparent counterexamples (§2). I then argued that the apparent counterexamples cannot be explained away by appeal to a categorical view (§3). I conclude that *Counterfactual Entailment* is a true principle whose apparent counterexamples should be explained away by appeal to the suppositional view.<sup>13</sup>

University of Colorado at Boulder  
Department of Philosophy  
232 UCB, Hellems 167  
Boulder, Colorado 80309-0232  
USA  
David.Barnett@Colorado.edu

#### References

- Adams, E. W. 1965: 'A Logic of Conditionals'. *Inquiry*, 8, pp. 166-97.  
———. 1966: 'Probability and the Logic of Conditionals'. In *Aspects of Inductive Logic*, edited by J. Hintikka and P. Suppes. Amsterdam: North Holland, 1966, pp. 256-316.  
———. 1975: *The Logic of Conditionals*. Dordrecht: Reidel.  
———. 1998: *A Primer of Probability Logic*. Stanford, CA: CSLI Publications.  
Barnett, D. 2006: 'Zif is If'. *Mind*, 115, pp. 519-566.  
———. 2009: 'The Myth of the Categorical Counterfactual'. *Philosophical Studies*, 144, pp. 281-296.  
———. 2010: 'Zif Would Have Been If: A Suppositional View of Counterfactuals'. *Noûs*, 44, pp. 269-304.  
Barker, S. 1999: 'Counterfactuals, Probabilistic Counterfactuals, and Causation'. *Mind*, 108, pp. 427-469.  
Dummett, M. 1978: *Truth and Other Enigmas*. London: Duckworth.  
Edgington, D. 1995: 'On Conditionals'. *Mind*, 104, pp. 235-330.

---

<sup>12</sup> For deeper developments of the suppositional view and further arguments in its favor, see Barnett 2006, 2009, 2010, Edgington 1995, and Mackie 1973.

<sup>13</sup> For helpful comments I am grateful to Yuval Avnir, Mike Huemer, Peter Kung, John Morrison, and Adam Pautz. I also received helpful feedback during presentations of drafts of this paper to philosophy departments at the University of Colorado (2008), University of Wyoming (2009), UNC-Chapel Hill (2009), and Oxford University (2010), and to the Graduate Philosophy Conference at the University of Texas at Austin (2009), the Midwest Undergraduate Philosophy Conference at Creighton University (2011), and the Aristotelian Society (2011).

- Goodman, N. 1947: 'The Problem of Counterfactual Conditionals'. *Journal of Philosophy*, 44, pp. 113-28.
- Harper, W. L., Stalnaker, R., and Pearce, C. T. (eds.) 1981: *Ifs*. Dordrecht: Reidel.
- Hiddleston, E. 2005: 'A Causal Theory of Counterfactuals'. *Noûs*, 39, pp. 632-657.
- Lewis, D. 1973: *Counterfactuals*. Cambridge, Mass: Harvard University Press.
- Mackie, J. 1973: *Truth, Probability and Paradox*. Oxford: Clarendon Press.
- McDermott, M. 1996: 'On the Truth Conditions of Certain "If"-Sentences'. *Philosophical Review*, 105, pp. 1-37.
- Pollock, J. L. 1976: *Subjunctive Reasoning*. Boston: Reidel.
- Quine, W. V. 1950: *Methods of Logic*. New York: Holt, Rinehart, and Winston.
- Ryle, G. 1950: "If", "so" and "because". In *Philosophical Analysis*, edited by M. Black. Englewood Cliffs: Prentice-Hall.
- Stalnaker, R. 1968: 'A Theory of Conditionals'. In *Studies in Logical Theory, American Philosophical Quarterly Monograph* Volume 2, pp. 98-112.
- . 1981: 'A Defense of Conditional Excluded Middle'. In W. Harper, G. Pearce, and R. Stalnaker, 1981, pp. 87-104.
- Williamson, T. 1996: 'Knowing and Asserting'. *Philosophical Review*, 105, pp. 489-523.
- Woods, M. 1997: *Conditionals*. Oxford: Clarendon Press.