Morgenbesser's Coin

Yael Loewenstein
yrloewen@central.uh.edu
University of Houston

Abstract

Before a fair, indeterministic coin is tossed, Lucky, who is causally isolated from the coin-tossing mechanism, declines to bet on heads. The coin lands heads. The consensus is that the following counterfactual is true:

M: If Lucky had bet heads, he would have won the bet.

It is also widely believed that to rule M true, any plausible semantics for counterfactuals must invoke causal independence. But if that's so, the hope of giving a reductive analysis of causation in terms of counterfactuals is undermined. Here I argue that there is compelling reason to question the assumption that M is true.

Keywords: Morgenbesser's Coin; Counterfactuals; Semantics of Counterfactuals; Causal Independence; David Lewis; Counterfactual Analysis of Causation

I. INTRODUCTION

Imagine a fair, indeterministic coin-tossing machine that fires a single photon toward two slits whenever the machine's activation button is pressed. Which slit the photon enters is genuinely random: a complete description of the world at any moment prior to the event, together with the laws of nature, does not fix the outcome. If the photon goes through the left slit, it causes the coin to land heads. If it goes through the right slit, it causes the coin to land tails. Since it is a fair coin tossing machine, each photon has a 0.5 chance of going into the left slit and a 0.5 chance of going into the right slit.

Let us imagine that a "coin tosser" named Susan is about to press the activation button. Just prior to her pressing it, Lucky is given the chance to bet that the coin will land heads. Instead, Lucky bets tails. We stipulate that Lucky's bet has no causal impact on Susan's button

pressing. Susan presses the button and the coin lands heads. This is the setup for an important counterfactual called *Morgenbesser's Coin:*[1]

(M): If Lucky had bet heads he would have won.

Most people immediately judge (M) to be true. Call the intuition that (M) is true the "ordinary intuition". Here I argue that the ordinary intuition is wrong and that in fact, (M) is false. While I do not claim to offer a knockdown proof of (M)'s falsity, my aim is to put the counterfactual's purported truth into serious question. The truthvalue of (M) is important, not only for the semantics of counterfactuals, but also for questions in metaphysics and the metaphysics of science, among other areas. What follows is just a sample of some of the reasons that (M)'s truthvalue matters.

First, Morgenbesser's Coin has had, and continues to have, an extraordinary influence on the literature on counterfactuals. I will cite some examples of its impact. The Lewisian (1973, 1979) truth conditions for counterfactuals are below:

> A counterfactual "If it were that A, then it would be that C" is (non-vacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than is any world where A is true but C false. (Lewis 1979, 465)

The similarity ordering is given by the following similarity weighting (paraphrased from Lewis 1979, 472):

1. It is of the first importance to avoid big miracles.
2. It is of the second importance to maximize the region of perfect match.
3. It is of the third importance to avoid small miracles.
4. It is of little or no importance to maximize the region of imperfect match.

[1] Morgenbesser's Coin was named after Sidney Morgenbesser (discussed in Michael Slote 1978).

This ordering leaves open whether imperfect match of particular fact should count for nothing, or whether it should have relatively little weight. This reflects Lewis's own uncertainty:

> It is a good question whether approximate similarities of particular fact should have little weight or none. Different cases come out differently, and I would like to know why. Tichy and Jackson give cases which appear to come out right under [the analysis shown above] only if approximate similarities count for nothing; but Morgenbesser has given a case, reported in Slote ([1978]), which appears to go the other way. (Lewis 1979, 472)

Despite the fact that it was Morgenbesser's Coin that apparently motivated Lewis's inclusion of the fourth criterion (the first three, by themselves, rule the counterfactual *false*[2]), the similarity ordering given above is nevertheless still unable to rule (M) true, even with the inclusion of (4). That is because, although a (Lucky-bets-heads-and-coin-lands-heads)-world, call it w1, preserves the outcome of the coin toss—and thus, one aspect of imperfect match—a (Lucky-bets-heads-and-coin-lands-tails)-world, w2, preserves a different aspect of imperfect match: match in the outcome of the bet (i.e., in w2, like in the actual world, Lucky *loses*). Since, as Jonathan Schaffer points out, "either [match in coin toss outcome or match in bet outcome] might have the wider ramifications – for instance, either might inspire Nixon to press the button [resulting in nuclear war]" (Schaffer 2004, 303), we can easily revise our background story so that (1)-(4) rules (M) as clearly *false* (if w1 and w2 turn out to be equidistant from the actual world, (M) is also false).

In response to this apparent problem, many proponents of the traditional possible-worlds framework have attempted to modify Lewis's similarity metric so that it can rule (M)

[2] To rule (M) true, the post-antecedent fact that the coin actually lands heads must count for similarity. The majority of other post-antecedent facts must *not* count for similarity, however, since these facts must vary along with the changes brought about by the antecedent-event in order to satisfy criterion (1). This means that the coin landing heads is part of a region of imperfect match, so regions of imperfect match must count for similarity for (M) to come out true.

true, regardless of the ramifications of the toss's outcome. The consensus seems to be that it should be modified in the following way: all and only (relevant) facts that are *causally independent* of the event referred to by the counterfactual's antecedent should be counted for comparative world similarity.[3] Call this the *causal independence thesis*, or *CIT*. If all and only the relevant facts causally independent of the antecedent count toward similarity (that is, if CIT is true), then given that the outcome of the toss is relevant in the context, (M) is true.[4] That is because although the outcome of the toss is both causally and probabilistically independent of Lucky's bet, the outcome of his bet—i.e., whether he wins or loses—is not. On this proposal, Lucky-bets-heads-coin-lands-heads worlds are closer than Lucky-bets-heads-coin-lands-tails worlds, because the preservation of the toss outcome counts towards similarity, whereas the preservation of the betting outcome does not.

Incorporating CIT to get (M), and counterfactuals like it, to come out true is no trivial revision to the semantics for counterfactuals. It has serious implications for ongoing disputes in metaphysics and the philosophy of mind, among other areas.[5] Perhaps most significantly, it

---

[3] See, e.g., Bennett (2003), Schaffer (2004) and Edgington (2004), although Schaffer seems to understand causal independence as only a *necessary* condition for facts counting toward similarity. (Thanks to Alan Hájek, here.)

[4] Noordhof (2004) maintains that we can get (M) to be ruled true by only counting toward similarity facts *probabilistically* independent of the antecedent. The problem, as Schaffer (2004) has pointed out, is that although whether the coin lands heads or tails is probabilistically independent of Susan's toss, it seems that the outcome of the bet (that is, whether Lucky wins or loses) is *also* probabilistically independent of the toss - and yet we don't want to count that as relevant for similarity. Noordhof (2005) has a way around this, although it requires accepting other of his controversial commitments. Nonetheless, each of the arguments I give here works just as well against the view that only facts probabilistically independent of the antecedent should count toward similarity.

[5] One example of how (M)'s truthvalue has important implications for disputes in the philosophy of mind is its implications for the *causal exclusion problem*: roughly the problem for dualism that if mental properties are not physical, the physical causes of behavioral effects seem to exclude the possibility that there are mental causes for those effects. If (M) is false, then the standard account of causal overdetermination as applied to indeterministic contexts cannot be right. According to the standard account, an event, e, is overdetermined by events c1 and c2 only if, had c1 happened without c2, or c2 happened without c1, e would still have happened. This counterfactual test for overdetermination is usually presupposed in discussions of the causal exclusion problem. For example, Karen Bennett (2003) has influentially argued that there is no serious exclusion problem because the mental and physical fail to meet the above stated counterfactual test. But if (M) is false, the above test is *not* the right test for causal overdetermination, since c1 and c2 can intuitively overdetermine e even if, had c1 not have happened, c2 might have failed to cause e (and so e might not have happened). To see this, consider an example: assassins Alice's and Bob's gun shots can intuitively overdetermine a target's death if both hit him in the heart simultaneously, despite the fact that, if counterfactuals like (M) are false, and supposing that (i) the two shootings

means understanding counterfactuals *in terms of causation*, which rules out giving a reductive analysis of causation in terms of counterfactuals. Indeed, invoking CIT in one's analysis of counterfactuals threatens to undermine the widely-held, broadly Humean conception of causation according to which causation can be reduced, in some way, to counterfactual dependence.

The assumed truth of (M) has had, and continues to have, a significant impact on the literature on counterfactuals in other ways as well. Some theorists have cited an ability to rule (M) true as evidence for alternative semantic accounts that depart from the Lewis-Stalnaker picture entirely. Hiddleston (2005), for example, has appealed to the apparent truth of (M) as evidence for his causal-model based theory of counterfactuals. Indeed, if it turns out that (M) is actually false, Morgenbesser's coin will provide us with a counterexample to causal-modeling accounts of this sort. And Khoo (2017) appeals to its truth in defense of his historical modality theory of counterfactuals.

Given how much hangs on the truthvalue of (M), it has been a mistake, I think, to simply take the ordinary intuition for granted. Especially since, as I will now argue, there are reasons to question the assumption that Morgenbesser counterfactuals are true.

II. QUESTIONING (M)'S TRUTH

II1. Two Ways to Think About (M)

There are two different ways to think about Morgenbesser's Coin. Thought about one way it seems evident that the counterfactual is true. Reasoning about the counterfactual in a second way, however, leads to the opposite conclusion. I suspect that those who immediately judge (M) to be true do so because they are thinking about the counterfactual in the first way. The

---

are causally independent and (ii) the laws are indeterministic, had Alice's bullet not hit the target then Bob's may not have, either.

problem is that, as we shall see, the first way is unmotivated in an indeterministic context. The first way to reason about (M) is as follows. Given that the coin actually lands heads, and that, had Lucky bet heads, his bet would have had no impact on the toss, his bet would have made no difference to the *outcome* of the toss, either. Therefore, the coin would have (*still*) landed heads. Arif Ahmed explicitly reasons in this way in defense of the ordinary intuition about (M) (Arif 2011, 80) [6]:

> (1) If C makes no difference to an actual event E then E would still have occurred even if C had not (premise).
> (2) C makes no difference to any actual events to which it is causally irrelevant (premise).
> (3) [Lucky's not betting heads] is causally irrelevant to the [outcome of the toss] (premise).
> (4) Therefore [M] is true.

Here, on the other hand, is the second way to reason about (M). Consider how Alexander Pruss (2003, 624) describes our "ordinary thinking" about counterfactuals:

> In the case of our ordinary thinking about counterfactuals, it is natural to locate, with some vagueness, the first event in respect of which the counterfactual world is supposed to diverge from the actual world, and then to consider how the divergence causally propagates as a result of this event.

Counterfactual semantic models generally aim to capture something approximating how we ordinarily understand counterfactual assertions. And we ordinarily understand a counterfactual assertion as asserting something like the following: if the antecedent (which in most cases is actually false) were true, and if the world prior to the antecedent were otherwise approximately the same (with only minor differences required to make the antecedent true),

[6] As far as I know, Ahmed's argument is the only argument made in defense of the ordinary intuition about (M). In all other cases, that (M) is true is simply taken for granted.

then the consequent would follow. We make the required changes *prior* to the time of the antecedent's obtainment, and let things unfold, as they will, from there. If the consequent obtains, the counterfactual is true. If the consequent does not obtain, the counterfactual is false. Let's start with this rough and ready picture and see what happens when we use it to evaluate (M). First, we "go back" to a time just before the time of the antecedent, when the (minimal-possible) changes need to be made for the antecedent to obtain: in this case, for Lucky to decide to bet heads. Call the moment of the first required change the *fork*. We then let the subsequent events unfold.

In a *deterministic* world, since Lucky's bet plays no causal role in the outcome of the coin toss, the coin would still land heads (we assume that none of the minor changes required to make Lucky decide to bet heads would themselves have a causal impact on Susan's button pressing, either). Since the outcome of the coin toss is stipulated to be indeterministic, however, if we let the sequence of events play out from the point at which Lucky bets heads, there is no guarantee that the coin will still land heads. It could land either heads or tails, despite not being influenced by Lucky's bet.

We can understand the second way to reason about (M), then, as follows. The change that results in Lucky betting heads takes us on a different post-fork path – or, in world talk, it takes us to a different world or worlds. Had Lucky bet heads the world would not have been the actual world ('actual world' should be understood as a rigid designator, here). And, since the outcome of the toss is indeterministic, at different worlds the coin could land either heads or tails, despite the coin actually landing heads. Since the coin might land tails at the relevant world (or worlds) at which Lucky bets heads, (M) is false.

This way of reasoning gives us a way out of Ahmed's argument. Ahmed's first premise says that if C makes no difference to an actual event E, then E would still have occurred even

if C had not. Substituting in the actual fact that Lucky did *not* bet heads for C, and the actual fact that the coin landed heads for E, premise (1) says that had C not occurred—that is, had Lucky *bet* heads—the coin would have (still) landed heads. On our alternative picture, this premise is false. E can fail to occur following C not occurring, despite C not making a causal difference to E. That is because the change which results in C not occurring, puts us at a different world. And at a different world, the coin could land heads or tails, regardless of what it lands at the actual world. C need not be causally connected to E for a change in C to correspond with a possible change in E.

II2. Intuitions About (M)

So which way of thinking about (M) is right? It seems that Ahmed has ordinary intuition on his side. Each premise of his argument, including premise (1), is intuitively appealing at first sight. And of course, many have the intuition that (M) is true. But can we trust our immediate intuitions about *a genuinely indeterministic* coin tossing process, or could it be that when we form these intuitions we fail to take adequate account of the role that indeterminism plays?[7] We are not used to thinking about genuinely random processes, after all, especially ones whose outcomes, e.g., whether the coin lands heads or tails, are *not* causally affected by macro-level details (like how the coin is tossed). Recall that it has been stipulated that the coin has a 0.5 chance of landing heads, regardless of the particular details of the token button pressing. This makes the scenario different from just about any scenario we come across in ordinary experience: even if the world is indeterministic, most outcomes still have many conspicuous causal influences. Whether the basketball goes through the hoop or not may not be causally *determined*, for instance, but the chance of each outcome is still causally

[7] Phillips (2007), (2011) makes a similar point, though he fleshes it out differently.

affected by the details of how the ball is thrown.

That Ahmed's first premise seems *prima facie* plausible is unsurprising, whether or not it is actually true. We know that Ahmed's first premise is true in a *deterministic* context: if C makes no causal difference to E and E is the result of a deterministic sequence, then E would still have occurred even if C had not. What's more, it will presumably *seem* true at a world where indeterminism does not play a highly conspicuous role. Say that the laws at the actual world are indeterministic. Say, furthermore, that given the details of precisely how James Harden shoots the basketball in a particular token shot attempt (for instance), the shot has a 99.9% chance of going in. It would take something of a quantum or statistical mechanical "fluke" for the ball not to go in, given the position, angle, and force exerted. Harden shoots and scores. In that case, holding fixed all causally relevant details, it is still exceedingly likely that Harden would have made the shot under the counterfactual supposition that some causally *irrelevant* detail or details were different: the shot would have still had a 99.9% chance of going in, even if Ahmed's first premise is false (of course, if the premise is true, the chance of Harden making it is 1). The point is that if this is the sort of inconspicuous role that indeterminism generally plays in our ordinary experience – i.e., most events have chances of close enough to 1 (given all relevant facts) for us to not be able to recognize an indeterministic component – then we should expect Ahmed's first premise to ring true, even if it is false.[8] And in that case, we should *not* take its intuitive plausibility as compelling evidence for its truth.

II3. No Relevant Difference Between Causally Dependent and Independent Facts Under Indeterminism

I now argue that, under indeterminism, there is *no good reason* to justify ignoring (only) facts causally dependent on the antecedent in the determination of world-similarity in the

[8] Of course, we don't usually know all relevant facts anyway, so the chances generally don't even need to be very close to 1 for us to not be aware of the indeterministic influence, if there.

evaluation of (M). And if that is right then ignoring such facts amounts to an ad hoc 'fix', no better than, say, ignoring all facts known by the Queen of England if doing so gets (M) to be true. There is nothing significant about the causal dependence/independence distinction which justifies treating causally dependent facts differently from causally independent ones. We can see this if we consider why causally dependent facts might be thought to be special (and why they *are* special under determinism). As far as I know, Jonathan Schaffer is the only one to explicitly say why the causal dependence/independence distinction should matter for world-similarity. Schaffer writes the following:

> Here is one way to express [the idea of invoking causal independence]: only match among those facts *causally independent of the antecedent* should count towards similarity. After all, if outcome o causally depends on p or ~p , then o should be expected to *vary* with p or ~p – its varying should hardly count for *diss*imilarity." (2004, 305, his emphasis)

In general, if some effect, E, causally depends on some cause, C, then – barring cases of causal overdetermination – it is expected that E will vary with C at the closest C and ~C worlds. C worlds at which E does not obtain, *or* ~C worlds at which E does, are, if anything, *further away* from the actual world, than C&E or ~C&~E worlds: E's failure to vary with C at a world suggests that there isn't the same relation of causal dependence at that world.

For this reason, it makes sense to think that we should only count for similarity that which is causally independent of the antecedent and allow that which is causally dependent on the antecedent to vary with the antecedent as it will. But notice that this motivation for CIT does not extend equally well to all cases. In particular, it does not extend to worlds at which the probability of the effect is exactly the same, given the cause. For in that case, there is no reason to think that varying the cause should necessarily result in a variation in the effect.

To see this, consider the following counterfactual (discussed, for different purposes, by Bennett (2003)):

(B) If Lucky had tossed the coin, it would have landed heads.

At the actual world, $w_@$, Susan presses the button to toss the coin. To evaluate (B) we must look to the nearest worlds at which Lucky presses the button to toss the coin, instead. Call one of the nearest worlds to $w_@$ at which Lucky presses the button and the coin lands heads, *w1*, and one of the nearest worlds to $w_@$ at which Lucky presses the button and the coin lands tails, *w2*.[9] Since whether it is Susan or Lucky who tosses the coin makes no difference to the chance of each outcome, there is no longer the same justification for thinking that the outcome being the same at $w_@$ and w_1 *shouldn't* make the worlds more similar to one another than $w_@$ is to w_2.[10] That is, if the only reason to think that match in outcome does *not* matter for similarity is that for any given cause and effect, we expect the effect to vary when we vary the cause, then why *shouldn't* match in outcome count here? Aren't two worlds at which the coin is tossed and lands heads *prima facie* more similar to one another than to a third at which it lands tails? We could try to appeal to context to say why the outcome of the toss is not contextually relevant to similarity, but that is not a promising route: coin toss outcome is precisely what is relevant. I submit that the fact that the outcome is the same at both worlds should now, arguably, count in favor of their similarity if imperfect match is relevant for similarity at all (this is a good reason to deny that imperfect match is relevant for similarity at all![11]).

So it remains mysterious why we should think that *even in an indeterministic context* there is something significant about facts that are causally dependent on the antecedent which can justify disregarding (just) those facts in our assessment of world similarity. We can make a

[9] For ease of exposition, I speak as though the 'Limit Assumption', the assumption that there is always a set of most similar antecedent-worlds, holds. Many deny this assumption, but nothing hangs on it here.
[10] I thank Carolina Sartorio for help with this way of putting the point.
[11] If we deny that imperfect match is relevant to similarity (given a similarity semantics), we evade this difficulty. On Lewis's picture denying that imperfect match is relevant means removing his fourth criterion of similarity.

related point in a different way, using (B). I now argue that CIT makes the *wrong* ruling for (B) and other counterfactuals like it, and that this provides a good reason to reject CIT. If it turns out that CIT is in fact wrong, then the best (and indeed, so far, *only*) proposals for ruling (M) true will have been refuted.[12] This would not bode well for the ordinary intuition about (M).

III. CIT's RULING ON (B)

To temper expectations, I should reiterate that I do not profess to be offering any kind of *proof*, either against CIT or for the falsity of (M). My aim is merely to raise some degree of skepticism about the widely-held assumption that (M) is true. Recall (B):

(B) If Lucky had tossed the coin, it would have landed heads.

I think it will be agreed that (B) is false. If Lucky had been the one to toss the coin, it may have come up either heads or tails. Indeed, we need not rely only on our intuitions for evaluating this counterfactual. Since which slit the photon enters is stipulated to be indeterministic, even fixing all the facts – e.g., when the button is pressed, how the button is pressed, and so on – is, by definition, insufficient to fix the outcome. And here we need *not* even fix these facts. For my purposes all that needs to be stipulated is that the outcome is genuinely random, and that the coin has a 0.5 chance of landing heads and a 0.5 chance of landing tails regardless of the particular details of when and how the button is pressed. We are free to assume, for instance, that it is implicit in the context that had Lucky been the one to press the button he would have done so differently than Susan did: for example, he probably would *not* have pressed it at precisely the same moment that she did. We may even imagine, if it is helpful, that he would have pressed it significantly earlier than she did. If the button

[12] For simplicity I focus exclusively on the causal independence thesis, although my argument can be used against the probabilistic version of the thesis just as well. The probabilistic version says that all facts probabilistically independent of the antecedent ought to be held fixed when evaluating a forward-tracking counterfactual.

pressing is done by a different person, at a different time and in a different way, then it is, surely, a distinct button pressing. And we make the reverse gambler's fallacy if we think that because the chancy coin landed heads in one particular trial, it is guaranteed to do so in a different (actual or counterfactual) one.

Does CIT correctly predict that (B) is false? To answer this question, we should ask the following one: is the fact that Lucky does not toss the coin at the actual world causally relevant to the coin landing heads? If it is, then CIT says we should not hold fixed that the coin lands heads, and (B) comes out false as it should. If, however, the coin landing heads is causally independent of Lucky not tossing the coin, then CIT prescribes holding the outcome of the coin toss fixed, and (B) comes out true. But how to decide if the two are causally independent? On the one hand, had Lucky been the one to toss the coin it *could* have landed tails. There is a 0.5 chance that it would have. But of course, it also could have still landed heads. Does the fact that the outcome could have been different had Lucky tossed the coin – even if it is just as likely that it would not have been different – indicate that the outcome is causally dependent on Lucky *not* tossing the coin? I think the answer is no: the outcome is causally *independent* of Lucky not tossing the coin. The reason is simple. Once again, since the machine is indeterministic and since there is stipulated to be a 50-50 chance that the photon will enter the left (or right) slit given that the button is successfully pressed at all, *which* slit the photon enters is causally independent not only of who presses the activation button, but also of when it gets pressed, how it gets pressed, and so on. Because the outcome is genuinely indeterminate - and crucially, because there is no influencing the chance that the coin will land heads rather than tails or vice versa - whether the coin lands heads rather than tails is causally independent of the entire conjunction of facts that together constitute every describable aspect of the button pressing.

Notice that to make this assertion I do not need to assume that causation requires determination. I am relying on a much weaker, seemingly unobjectionable assumption that for some event to be caused, its occurrence should not be *wholly* up to chance. What do I mean by *wholly* up to chance? Compare the coin toss scenario to a scenario in which I toss a rock at a window with the intention of breaking it. If the world is indeterministic, then it may be that a complete description of the rock toss, the window, the conditions outside, and so on, conjoined with the laws of nature, is insufficient to determine whether or not the window breaks: that is, there may be an element of chance involved. But even if there is an element of chance, the outcome of the rock toss is clearly not wholly up to chance. Whether I succeed at shattering the window depends not only on chance, but also on how I toss the rock.

This is why it is important in the coin toss scenario that we consider an idealized case in which the details of the button-pressing make no difference whatsoever to the chance of each outcome. And it is standard to think that in ordinary, benign cases in which there is not causal overdetermination and where probabilistic preemption does not occur, given two actual events C and E, *C must make a difference to the probability of E if C is a cause of E.*[13] But, Lucky not pressing the button makes no difference to the probability that the coin lands heads.

Why might one think that Lucky not tossing the coin *does* makes a causal difference to the outcome of the toss? There is one sense in which who tosses the coin is causally relevant to the outcome; it is just not the sense that matters for the evaluation of the counterfactual. There is an important distinction between causing the coin to fly into the air and so to land face-up at all, and thereby causing it to land heads (if it happens to do so) in virtue of *that*, vs.

[13] Probabilistic preemption occurs when there are two potential causes, C1 and C2, of some effect, E. Both causes individually raise the probability of E if taken alone, however the probability of E is higher conditional on C2 than it is on C1. If it is the case that, say, C2 occurs if and only if C1 does not, then C1 lowers the probability of E even if C1 is the actual cause of E.

causing the coin to land heads *rather than* tails, or tails *rather than* heads, given that it is tossed. Suppose that I push the button and the coin lands heads. I caused it to land heads in the first sense because I caused it to land something or other. In addition, it happened to land heads rather than tails (for a-causal reasons) and, because (a) I caused it to land something or other and (b) it happened to land heads rather than tails, I caused it to land heads. Had it landed tails rather than heads (for a-causal reasons) I would have caused it to land tails, again in virtue of causing it to land something or other.[14]

It is the second component, the acausal part, that interests me here. That is because it is this part, I will argue, that is relevant to the evaluation of (B) and (M). But first, to bring out the distinction more clearly, it will be helpful to compare our scenario to another scenario involving probabilistic causation. Imagine that a terrorist is considering setting an indeterministic time bomb which, if set, has a 0.5 chance of detonating regardless of the details of how it is set. In this scenario there are at least the following three contextually salient possibilities:

(i) The terrorist successfully sets the bomb and it detonates.
(ii) The terrorist successfully sets the bomb and it does not detonate.
(iii) The terrorist does not successfully set the bomb (and it does not detonate).

Suppose the terrorist decides to set the bomb and it detonates. By setting the bomb, the terrorist eliminates possibility (iii). For the purposes of this example, let us adopt a probabilistic understanding of indeterministic causation according to which *to cause* is

[14] The contrastivist about causation, who holds that causation is not a simple two-place relation between cause and effect, but rather a three- or four-place relation involving either contrastive causes or contrastive effects or both, can handle this distinction well. Nonetheless, I don't want to commit myself to the view that all causation is contrastive. I maintain only that a meaningful distinction can be made between something being causally relevant to the coin landing heads in virtue of being causally relevant to it landing face-up in some way or other, versus being causally relevant to whether the fair and indeterministic coin lands heads or tails given that it is tossed. If such a distinction, which seems perfectly intelligible, is deemed incoherent or cannot be meaningfully expressed by a non-contrastivist account, then I say so much the worse for the non-contrastivist account.

understood as something like *to make more likely to happen.*[15] The terrorist's action is causally relevant to the detonation of the bomb in the following way: by setting the bomb, she eliminates possibility (iii), and thereby raises the probability of (i) to 0.5. Had (i) obtained, the terrorist would be causally responsible for (i) in virtue of having eliminated possibility (iii) (or alternatively, if it is preferred, we can instead say that the terrorist would be causally responsible for (i) in virtue of bringing about the possibility that (i) could obtain). But here is the crucial point: the terrorist is *not* causally responsible for which of the remaining two possibilities—i.e., either (i) or (ii)—obtains if she sets the bomb. Given that she sets it, *whether* it goes off or not is entirely up to chance. (Note that the scenario could be revised so that someone's action *is* causally relevant to which of (i) or (ii) obtains if the bomb is set. If a different terrorist, say, tampers with the bomb to make it such that if it is set then (i) is more likely to occur than (ii), then that second terrorist's action plausibly would be causally relevant to (i) occurring instead of (ii) if (i) actually occurs.)

I pause to flag that the distinction I am making requires that I depart a bit from the standard probability-raising model of probabilistic causation in order to capture the idea that nothing is causally responsible for whether (i) or (ii) obtains given that the terrorist sets the bomb. In its most naïve form, the standard probability-raising model for probabilistic causation says that some event C is a cause of (indeterministic) outcome E just in case the conditional probability of E given C is higher than the conditional probability of E given –C.[16]

[15] I choose this conception of indeterministic causation for simplicity and because it is, as far as I can tell, the most widely held account. Not much should be made of this choice, however. My argument could be made just as well in terminology consistent with most alternative conceptions of indeterministic causation.

[16] No one accepts this naïve formulation as is, since it cannot distinguish between C being a genuine cause of E and C being an effect of, or sharing a common cause with, E. But the details regarding how *Standard* has been or should be modified are not important here.

Standard: C is a cause of E iff Prob (E|C) > Prob (E|− C)

Applied to the case at hand, if the terrorist sets the bomb and it detonates, then by *Standard*, the detonation is causally dependent on the terrorist setting the bomb in virtue of the probability that it detonates conditional on her setting it being higher than the probability that it detonates conditional on her not setting it. This is the sense in which the outcome *does* causally depend on the terrorist setting the bomb. But the standard model does not seem applicable to causal dependence as it relates to the question of whether the bomb detonates or not if set (i.e., the question of whether possibility (i) or (ii) obtains once possibility (iii) is eliminated). To be clear, I am *not* claiming that the standard model gets the *wrong* answer to this question; rather, my claim is that the standard model does not seem to have anything to say about it at all. So, to the extent that whether the indeterministic bomb detonates or not given that it is set does not causally depend on anything is a meaningful idea – and I cannot see any reason to deny that it is – it seems that we will need to capture it in a different way. How it ought to be done is not important for my purposes. What matters is only that a distinction can be made: if the bomb detonates then the bomb detonation *does* causally depend on the terrorist setting the bomb in the ordinary probability-raising sense; however, it is also the case that in our idealized scenario, *whether* the bomb detonates or not *if set* is an acausal matter (i.e., it causally depends on nothing).

Let us return to (B). In most contexts in which (B) might be uttered, in the counterfactual scenario in which Lucky presses the activation button there are *two* salient alternatives. Either:

(iv) The coin lands heads. Or,

(v) The coin lands tails.

Unlike in the terrorist scenario in which there was the possibility that the terrorist might not

set the bomb at all, in the counterfactual scenario there is not the possibility that Lucky might not press the button. The antecedent tells us that in the counterfactual scenario Lucky presses the button. Since it is a given that he does, the truthvalue of (B) depends only on *which* of the two alternatives, (iv) or (v), obtains when he presses it. At the nearest worlds at which Lucky presses the button, does the coin land heads rather than tails or tails rather than heads? It is this which must be considered to evaluate the counterfactual – but it is just this, as we have seen, that is the *acausal* part. Whether the coin lands heads rather than tails or vice versa is causally independent of the describable aspects of the coin toss. Just as the terrorist in the first scenario would have no causal influence on which of (i) or (ii) obtains were she to set the bomb, Lucky pressing the button would have had no causal influence on which of (iv) or (v) obtained had the button been pressed. And if that is right, then in the respect relevant to the evaluation of the counterfactual, the coin landing heads is causally independent of Lucky not pressing the button. But in that case, CIT tells us to hold the actual fact that the coin lands heads rather than tails, fixed. If we hold fixed that the coin lands heads, however, then (B) comes out true. Since (B) is false, it can be concluded that CIT gives the wrong ruling for (B).

IV. SOME OBJECTIONS AND REPLIES

Could the best response be to deny that (B) is false? There are a number of reasons to not be tempted to this conclusion. Taking (B) to be true, and thus match in coin toss outcome to be relevant to similarity in the evaluation of (B), appears to commit us to accepting that all kinds of facts that clearly should not be relevant to similarity in a context actually are. Is match in outcome relevant if the coin would have been tossed at a different time? What if it would have been tossed by a different person at a different time? What if, had Susan not tossed the coin in the USA, Lucky would have traveled with it to Canada and tossed it there? Should we say that it would have certainly landed heads, since it landed heads when Susan tossed it, too?

Is there a principled reason to say that match in outcome shouldn't count in this case but that it should count in the case of (B)? If there is, I don't know what it could be. And of course, the problematic examples proliferate well beyond coin tosses.

A better option is to accept that (B) is false and reject CIT. But perhaps this is still too quick. Perhaps we merely need to take a more permissive understanding of the concept of "causal independence". Surely we can understand CIT in a way that will allow it to avoid ruling (B) true. And indeed, we can. When Susan presses the button and the coin lands heads, the event of the coin landing heads lies on the causal chain that "passes through" (so to speak) Susan pressing the button. Likewise, had Lucky been the one to press the button, the coin landing face-up one way or another would have lay on the causal chain that passed through Lucky pressing the button. Thus, we can get CIT to rule that (M) is true and (B) is false if we understand the pertinent notion of causal dependence to be something like the following: two distinct events share a causal dependence relation just in case both events *lie on the same causal chain*. Jonathan Bennett (2003) defends a semantics for counterfactuals according to which causal dependence is understood in this way, precisely so that his semantics can rule counterfactuals like (M) true and those like (B) false.

> Bennett Semantics (BS): Counterfactual A>C is true if C is true at all the nearest A-worlds that maximize match with the actual world over events that *lie on the same causal chain* as at the actual world.[17](cf. 2003, 235)

BS rules (M) true without having to give up the falsity of (B). (M) is true because Lucky betting heads does not lie on the causal chain going from Susan's button pressing to the coin landing heads. So, by BS, the fact that the coin lands heads is relevant to world similarity, and (M) comes out true. In contrast, BS does not rule B true. If Lucky had pressed the button rather

[17] I am paraphrasing Bennett's view, omitting details that are not relevant here.

than Susan, that would have initiated a distinct causal sequence. So, while Lucky not pressing the button does not affect the probability that the coin lands heads, it does make it such that the *sequence* leading to the coin landing something or other is different than it otherwise would be. Thus, by BS, we do *not* count the outcome of the button pressing as relevant for similarity, and (B) comes out false.

The problem with the Bennett-style reply is that revising CIT as he does in order to get (M) true but (B) false is unacceptably ad hoc for just the reasons we've seen: it presupposes that whether the coin toss outcome lies on some given causal chain is relevant to the evaluation of the counterfactual. But this is not the case, here. To think that it is, is to conflate two ideas that I have argued need to be distinguished: being causally relevant to the outcome of the toss in virtue of causing it to land something or other, and being causally relevant to whether the coin lands heads rather than tails or vice versa. While it makes sense to speak of coin tosses and their outcomes as being the result of causal chains, it does not make sense to speak this way about *whether* the photon enters the left or right slit (and thus, whether the coin lands heads or tails) if the button is pressed, at least not in the case of (M) or (B). Since the truthvalue of (B) depends only on whether the coin would land heads *rather than* tails if it were tossed – just as the truthvalue of (M) depends only on whether the coin would land heads *rather than* tails were Lucky to bet heads prior to Susan tossing the coin – and since the chances are 50-50 no matter what – facts about causal chains have no role to play. And if that is right, it is good evidence against a semantic analysis which, to get counterfactuals like (M) true but those like (B) false, relies on a distinction between events which do, versus those which do not, lie on the same causal chain as the antecedent event.[18]

[18] If we could trust our intuitions about (M), it might be a different story: Bennett's account gets (M) and (B) *intuitively* right, and – were there not compelling reasons to question our intuitions about (M) – maybe that would be enough, no matter how ad hoc it is. The problem is that, as we saw in section II2, we *cannot* trust our

V. CONCLUSION

I have argued that the assumption that counterfactuals like Morgenbesser's Coin are true should be questioned. There is compelling reason to deny that the causal dependence-independence distinction is relevant to the similarity ordering in the evaluation of counterfactuals like (M) and (B). Relatedly, if we apply the notion of causal independence in the way that makes sense in the context of evaluating these particular counterfactuals, CIT results in the wrong truthvalue for counterfactuals like (B). This gives us good reason to reject it.

These conclusions, if right, bear on Lewis's fourth criterion of similarity: "it is of little or no importance to maximize the region of imperfect match". To advocate counting for similarity *all* facts causally independent of the antecedent is to advocate for maximizing post-antecedent regions of imperfect match. Recall that Lewis expressed puzzlement over why maximizing regions of imperfect match inexplicably seems necessary for getting the correct truthvalues for some counterfactuals (in particular, those like Morgenbesser's coin) but results in the wrong ruling for others. If (M) and counterfactuals like it are false, there is no longer any reason to think that maximizing regions of imperfect match - or that holding fixed post-antecedent facts of any sort - is *ever* required. As we saw in section I, however, the importance of the falsity of CIT and (M) goes well beyond the implications for Lewis's similarity ordering.

Acknowledgements:

For very helpful discussion and/or written feedback, I am grateful to Arif Ahmed, Cameron Buckner, Juan Comesaña, Alan Hájek, Boris Kment, Luis Oliveira, Carolina Sartorio, Matt Schuler, Jason Turner, and a number of anonymous referees. I am especially grateful to Terry Horgan, my dissertation advisor, who read multiple drafts and whose objections helped me to improve the paper considerably. Thanks also to audiences at

intuitions about (M), and so the ad-hocness charge is all the more important. (I thank an anonymous referee for pressing me on this.)

Metaphysical Mayhem at Rutgers University (2016), the 2016 APA Central Division Meeting, the London School of Economics, The University of Arizona, and the University of Houston.

REFERENCES

Ahmed, A. (2011). "Out of the Closet", *Analysis* 77-85.

Barker, S. (1999). "Counterfactuals, Probabilistic Counterfactuals and Causation", *Mind* 108.

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

Bennett, K. (2003). "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It", *Nous*, Vol. 37, No. 3, pp. 471—497.

Edgington, D. (2004). "Counterfactuals and the benefit of hindsight". In *Causation and Counterfactuals*, eds P. Dowe and P. Noordhof, 12–27. London: Routledge.

Goodman, N. (1946). "The Problem of Counterfactual Conditionals", Reprinted in Goodman, *Fact, Fiction, and Forecast*, 4th ed. Cambridge: Harvard, 1979.

Hájek, A. (manuscript), *Most Counterfactuals are False*. ANU, monograph in progress.

Hiddleston, E. (2005). "A Causal Theory of Counterfactuals", *Nous* 39:4

Khoo, J. (2017). "Backtracking Counterfactuals Revisited", *Mind* 126: 503.

Lewis, D. (1973). *Counterfactuals*. Basil Blackwell Ltd., Malden, MA.

Lewis, D. (1979). "Counterfactual Dependence and Time's Arrow", *Nous*, Vol. 13, No. 4: 455-476.

Noordhof, P. (2004). "Prospects for a Counterfactual Theory of Causation", in *Cause and Chance: Causation in an Indeterministic World*, in P. Dowe & P. Noordhof (eds.), 188-201, Routledge.

Noordhof, P. (2005). "Morgenbesser's Coin, Counterfactuals and Independence", *Analysis* 65: 261-63.

Phillips, I. (2007). "Morgenbesser Cases and Closet Determinism", *Analysis* 67: 42–49.

Phillips, I. (2011). "Stuck in the Closet: A Reply to Ahmed", *Analysis*, Vol.71, No. 1: 91-96.

Pruss, A. (2003). "David Lewis's Counterfactual Arrow of Time", *Nous*: 606-637.

Schaffer, J. (2004). "Counterfactuals, causal independence and conceptual circularity", *Analysis* 64: 299–309.

Slote, M. (1978). "Time in Counterfactuals", *Philosophical Review* 87 (1): 3-27.