Aporia vol. 23 no. 2—2013

Counterfactuals and Causal Decision Theory

Kevin Dorst

In Newcomb's problem, an agent is faced with a choice between acts that are highly correlated with certain outcomes, but that do not cause those outcomes. Consideration of this type of case has led many decision theorists to reject evidential decision theory (wherein acts are evaluated by the evidence they provide about the world) in favor of some version of *causal* decision theory, wherein acts are evaluated by how likely they are to *bring about* desirable outcomes. One common version of causal decision theory uses the probabilities of *counterfactuals* to calculate expected utilities, since these are supposed to track causal relations.[1] In this paper I will examine whether this is a plausible way to develop causal decision theory, focusing on the version presented in Gibbard and Harper (1978). I will argue that it is not, for counterfactuals sometimes fail to track the relevant causal relations. I will first show that Gibbard and Harper's definition of "causal independence" (CI) in terms of counterfactuals is inconsistent with a plausible understanding of the nature of the correlations between acts and outcomes in Newcomb's problem, and then provide evidence that this inconsistency should lead us to reject their formulation of CI. Having done so, I will then argue that Gibbard and Harper's version of decision theory in fact *agrees* with evidential decision theory under a plausible interpretation

[1] A counterfactual is a sentence of the form, "If A were to happen, then B would happen." This approach to causal decision theory has its roots in Stalnaker 1972, though Gibbard and Harper 1978 give the first in-depth presentation of it.

Kevin is a senior majoring in philosophy and political science at Washington University in St. Louis. He is currently applying to Ph.D. programs in philosophy, where he hopes to further explore his interests in metaethics, formal epistemology, philosophical logic, and philosophy of language.

of Newcomb's problem. If I am right about this, the upshot is that theorists attracted to the "causal" approach developed by Gibbard and Harper must either (i) accept a theory that is much closer to evidential decision theory than it may first appear, or (ii) reject the explication of causal decision theory in terms of counterfactuals.

I. Newcomb's Problem and Counterfactuals

The standard version of *evidential* decision theory developed by Jeffrey (1983) calculates the expected utility of an act using the conditional probabilities of possible outcomes given the act in question. Thus if $S_1 \ldots S_n$ are the possible outcomes of an act A, $prob(S_i | A)$ is the probability of an outcome Si obtaining *conditional* on the act A being performed, and $D(S_i)$ is the desirability of each S_i, the evidential expected utility of A is

$$V(A) = \sum_i [prob(S_i | A) D(S_i)]$$

A basic form of evidential decision theory states that an agent ought to choose an act with maximal expected utility as calculated by $V(\bullet)$.[2]

Although this theory is elegant and plausible in many situations, it has been thought to run into difficulties in cases where acts are highly correlated with desirable outcomes that they do not bring about. The most famous example of this was presented in Nozick (1969):

> **Newcomb's Problem**: A subject is presented with two boxes: one closed and one open. In the open box he sees \$1000, and the closed box contains either \$1,000,000 or nothing, depending on whether a being called the Predictor has put the money in the box. The subject is given a choice between two options: take only the closed box, or take both boxes. If the Predictor has predicted that the subject would take only the one closed box, then it has already put \$1,000,000 in that box; if it has predicted that the subject would take both, then the closed box is empty. The Predictor is incredibly reliable at correctly predicting these choices—let's say it is correct 99% of the time. The subject knows all of this.

[2] More sophisticated version of evidential decision theory often impose further requirements, such as that an act be ratifiable (see Jeffrey 1983, §1.7), but we need not grapple with this notion here.

In this scenario, evidential decision theory recommends that the subject take only one box. After all, the probability that there is $1,000,000 in the covered box is 99%, conditional on his taking only one box, whereas this probability is only 1% conditional on him taking both boxes; therefore the evidential expected utility of one-boxing is much higher than that of two-boxing.

Though this verdict seems reasonable enough to some, others think it is quite obviously ridiculous. The problem, these theorists claim, is that taking only one box *in no way brings it about* that there is $1,000,000 in the covered box. How could it?—The money is already either in the box or not, and it is not as if the subject's action right now will cause it to appear or disappear.[3] Instead of an evidential decision theory—which measures the *auspiciousness* of acts—we need a *causal* decision theory that measures how *efficacious* each available act is at bringing about desired outcomes (Joyce 2007, 538).

It is not my intent to assess the merits of the underlying intuitions behind this line of thought, but rather to examine one way of developing it. In particular, I will examine a now standard way of formulating such a causal decision theory that was first presented by Gibbard and Harper (1978), in which causal relations are represented by calculating the probabilities of certain *counterfactuals*. That is, instead of determining the probability that outcome *S* will obtain *conditional* on act *A* being performed, we determine the probability that the sentence '*If I were to A, then S would obtain*' is true (153). The thought is that such propositions will track the causal efficacy that *A* has in bringing it about that *S*. If we represent this counterfactual as $A \Box\!\!\rightarrow S$, then Gibbard and Harper's proposal is that we calculate causal expected utility as follows (158):

$$U(A) = \sum_i [prob(A \Box\!\!\rightarrow S_i)D(S_i)]$$

Thus Gibbard and Harper's version of causal decision theory—which I will refer to as "counterfactual decision theory"—states that an agent should choose an available act with maximal expected utility as calculated by $U(\bullet)$.

Given this specification, Gibbard and Harper go on to argue that counterfactual decision theory gives the (putatively) correct verdict of two-boxing in Newcomb's problem. They first formulate a the following definition of "causal independence" (172):

[3] See Nozick 1969, 115–17; Stalnaker 1972; and Lewis 1979, 240 for developments of these considerations.

> **CI**: A state S is causally independent of the choice between acts A and B iff $(A\square\!\!\rightarrow S)\leftrightarrow(B\square\!\!\rightarrow S)$.

Thus (roughly) a state S is causally independent of two acts when the choice between them is irrelevant to bringing it about that S.[4] They then use this condition to derive the verdict of two-boxing. Let T_1 be the proposition that the subject takes one box and T_2 that he takes both boxes. We can break the relevant possible states of the world into whether the \$1,000,000 has been placed in the closed box or not, represented by the propositions M and ~M, respectively. Suppose that the desirabilities of receiving \$0, \$1000, \$1,000,000, and \$1,001,000 are 0, 10, 100, and 101, respectively. So the counterfactual expected utilities of T_1 and T_2 are:

$$U(T_1) = prob(T_1\square\!\!\rightarrow M)(100) + prob(T_1\square\!\!\rightarrow \sim M)(0)$$

$$U(T_2) = prob(T_2\square\!\!\rightarrow M)(101) + prob(T_2\square\!\!\rightarrow \sim M)(10).$$

Given this, Gibbard and Harper reason as follows. M is clearly causally independent of the choice between T_1 and T_2 (and hence satisfies CI), for the money has already either been placed in the box or not before this choice arises (180). Thus we have $(T_1\square\!\!\rightarrow M)\leftrightarrow(T_2\square\!\!\rightarrow M)$, and likewise $(T_1\square\!\!\rightarrow \sim M)\leftrightarrow(T_2\square\!\!\rightarrow \sim M)$; so $prob(T_1\square\!\!\rightarrow M) = prob(T_2\square\!\!\rightarrow M)$ and $prob(T_1\square\!\!\rightarrow \sim M) = prob(T_2\square\!\!\rightarrow \sim M)$. And if that is correct, then no matter what value is assigned to these probability functions, $U(T_2)>U(T_1)$, and therefore counterfactual decision theory recommends two-boxing (181).

However, I will contest this reasoning and argue that counterfactual decision theory in fact recommends one-boxing once we examine Newcomb's problem more closely. The problem is that CI does not capture the form of causal independence of M that is guaranteed by the scenario, so in fact M is *not* "causally independent" as Gibbard and Harper define the term.

II. Modal Reliability and Causal Independence

I will now argue that on a plausible understanding of Newcomb's problem, Gibbard and Harper's contention that M satisfies CI with respect to the choice between T_1 and T_2 (i.e. their claim that M is "causally independent" of this choice) is actually inconsistent with the scenario. Label the

[4] As I will be criticizing this formulation of causal independence momentarily, I should note that Gibbard and Harper do not offer an explicit defense of this formulation—they merely define it as such.

propositions that the Predictor predicts one-boxing and that he predicts two-boxing as *P1* and *P2*, respectively. We are informed that the Predictor is very *reliable*, understood at the very least as implying that $prob(P1 \mid T_1)$ and $prob(P2 \mid T_2)$ are both very close to 1, for this is required to make it so that evidential decision theory recommends one-boxing.[5] But beyond this, we are not told what it is that makes these conditional probabilities so high. Prima facie, one plausible proposal is that the Predictor is very *modally*[6] reliable at correctly guessing the subject's choice: in the vast majority of nearby possible worlds in which the subject faces this choice, the Predictor predicts correctly. Call this the *modal interpretation* of Newcomb's problem.[7] As it turns out, under the modal interpretation of the scenario, Gibbard and Harper's claim that M satisfies CI with respect to the choice between T_1 and T_2 is false.[8]

Before arguing for the inconsistency, let me briefly stipulate that for the rest of the paper I will be using a simplified version of the standard Lewisian semantics for counterfactuals. Thus $A \mathrel{\Box\!\!\rightarrow} B$ is true iff either (1) no *A*-world is modally accessible from the actual world (the trivial case), or (2) some centered sphere of worlds around the actual world contains an *A*-world, and the material conditional $A \supset B$ is true at every world in this sphere (1973, 16). I follow Gibbard and Harper (1978, 156–57) in simplifying Lewis's semantics by assuming there is always a unique *A*-world that is closest to the actual world. I do all of this simply to allow some level of definiteness—we do not need to go to the trouble of explicating the notions of "spheres of worlds" or "modal accessibility" beyond an intuitive level.)

Now to bring out the inconsistency. Suppose both that we take the modal interpretation of Newcomb's problem, and that *M* satisfies CI with respect to the choice between T_1 and T_2. Since our Predictor is modally reliable, in the vast majority of possible worlds (including all nearby worlds, let us say) his predictions will correctly track the subject's choice. Now T_1 is not an outlandish possibility at all, so clearly there are T_1-worlds that

[5] Gibbard and Harper freely admit this point (180–81).

[6] I do not think it is important what kind of modality is in question, as long as the worlds are ordered by a similarity relation. If one wants a determinate kind of modality, let us say we are talking about metaphysically possible worlds.

[7] Gibbard and Harper do not explicitly address the whether or not they understand the Predictor's reliability to have modal implications (see 181–82).

[8] I will return to defend the modal interpretation after I draw out this inconsistency.

are modally accessible from and nearby to the actual world (in fact the actual world may *be* a T_1-world). So go to the smallest sphere of worlds that contains any T_1-worlds; since these worlds are nearby, the Predictor will correctly predict this and therefore $T_1 \supset P1$ is true at those worlds. And in all other worlds in this sphere (wherein T_1 does *not* obtain), the material conditional is trivially true. Thus the truth-conditions for the counterfactual are satisfied, and the modal interpretation ensures that $T_1 \Box\!\!\rightarrow P1$ is true in the actual world. Precisely parallel remarks apply, *mutatis mutandis*, to $T_2 \Box\!\!\rightarrow P2$.

With this in hand, it is straightforward to show that the claim that M satisfies CI with respect to T_1 and T_2 is false. This amounts to the claim that $(T_1 \Box\!\!\rightarrow M) \leftrightarrow (T_2 \Box\!\!\rightarrow M)$ is true. Now note that $M \leftrightarrow P1$ is true in all the relevant worlds (since the money is put in the box iff the Predictor predicts one-boxing), and so by substitution of modal equivalents we have $(T_1 \Box\!\!\rightarrow P1) \leftrightarrow (T_2 \Box\!\!\rightarrow P1)$. Since the modal interpretation ensures that the left-hand side of this biconditional is true (from the previous paragraph) we have that $T_2 \Box\!\!\rightarrow P1$ is true. But from the modal interpretation we also have that $T_2 \Box\!\!\rightarrow P2$ is true, and since $P1 \leftrightarrow \sim P2$ is true in all the relevant worlds (the Predictor will always make one and never make both predictions), we now have that *both* $T_2 \Box\!\!\rightarrow P2$ *and* $T_2 \Box\!\!\rightarrow \sim P2$ are true. And since these are both non-vacuously true, this is a contradiction.[9] Thus it turns out that the modal interpretation of Newcomb's problem is inconsistent with the claim that M satisfies CI with respect to the choice between T_1 and T_2 – we must reject either the interpretation or the claim.

Which should we reject? Well first note that obviously M is "causally independent" of the subject's choice *in some sense* of the term—after all, the money is already in the box or not before the subject makes this choice. So in denying Gibbard and Harper's claim about CI we would merely be denying that they have the correct account of causal independence. I think this is precisely what we should do, for two main reasons: (1) there are good, independent grounds on which to accept the modal interpretation, and (2) there are good, independent grounds to reject the claim that CI correctly captures our intuitive notion of causal independence. I will develop each point in turn.

[9] Go to one of the T worlds in the smallest sphere, and the semantics say that both $T_2 \Box P_2$ and $T_2 \Box {\sim} P_2$ are true at that world. Since T_2 is true at that world, $P_2 \& {\sim} P_2$ is true at that world.

Why should we accept the modal interpretation? Put bluntly, because it seems that we can make no sense of the high conditional probability without this interpretation. We know that each of $prob(P_i \mid T_i)$ are very close to 1, so the subject is very confident that whatever he chooses, the Predictor will predict correctly. Could this be the case even when the subject believes that the Predictor is not modally reliable at predicting his choice? Suppose he thinks to himself, "In a substantial proportion of the nearby possible worlds in which I am faced with this choice, the Predictor predicts incorrectly." Can he really believe this and still be very confident that the Predictor will predict correctly *in the world he's in*? For all he knows, he's in any one of those many possible worlds he just mentioned, so if he thinks the Predictor often get's it wrong in them, isn't that *just to think* that the Predictor has a good chance of getting it wrong in whatever world he's actually in?[10] I admit that there's not much more to my case than intuition, but I think this intuition is fairly powerful. At the very least it is unclear how the subject's belief in the Predictor's reliability in the actual world could be significantly different from his belief in the Predictor's modal reliability in nearby worlds. Thus in the absence of some further argument, I submit that we should accept the modal interpretation as the most natural understanding of Newcomb's problem.

Turn now to the independent reasons for rejecting CI as the proper analysis of causal independence. We can bring this out by invoking the following principle of Forward Causation:

> **FC**: If event *B* causally depends on event *A*, then *B* does not occur before *A*.[11]

Now recall Gibbard and Harper's account of causal independence:

> **CI**: A state *S* is causally independent of the choice between actions *A* and *B* iff $(A \mathrel{\Box\!\!\rightarrow} S) \leftrightarrow (B \mathrel{\Box\!\!\rightarrow} S)$.

[10] I'm assuming here that the epistemically possible worlds relative to the subject largely overlap with the relevant kind of possible worlds referred to by the modal interpretation. Seeing as the relevant worlds will all be nearby (and thus very similar to the actual world and so likely not discernibly different to the subject), I do not think this assumption is problematic.

[11] Of course, such a broad principle is bound to be a little rough around the edges. To make it completely sound we would have to specify a reference frame, and perhaps rule out certain quantum events by fiat. But all I will need is that FC is true in general of the kinds of events that figure into human decision, which it clearly is.

The problem is that denying the right-hand side of the main biconditional of CI does *not* entail that S is causally dependent on the choice—for example, it is possible that a *non-causal determination relation* ensures that the equivalence on the right-hand side is false.

To see this, suppose that right now I face the choice between going to the zoo today or not doing so; call these options Z and $\sim Z$ respectively. And assume that this is the only chance I will ever have to go to the zoo (perhaps I will die tomorrow, and I have never done it before). Now consider the possible state of affairs:

G: One year ago, I was one year away from going to the zoo.

Note that this is a determinate state of affairs that either obtained or not at the specified time; it is precisely analogous to the fact that as I type this sentence I occupy a state of affairs in which I will soon type a period.[12] Now, if I were to decide to go to the zoo today, then G would have obtained a year ago; if I were to decide against going to the zoo today, then G would *not* have obtained a year ago. Therefore $Z \Box\!\!\rightarrow G$ and $\sim Z \Box\!\!\rightarrow \sim G$ are both true. And since we can assume that the second counterfactual is non-trivially true (there are accessible $\sim Z$-worlds), we have that $\sim Z \Box\!\!\rightarrow G$ is *false*. Since $Z \Box\!\!\rightarrow G$ is true, it turns out that $(Z \Box\!\!\rightarrow G) \leftrightarrow (\sim Z \Box\!\!\rightarrow G)$ is false. Thus by CI it turns out that G is causally *de*pendent on my choice between Z and $\sim Z$. However, this result violates FC: G is a state of affairs that either obtained or not a year before my choice, so it *cannot* causally depend on my choice without allowing reverse causation. Therefore we must reject the claim that the counterfactual equivalence in CI tracks *causal* dependency. The problem is that there is a non-causal determination relation between G and the choice between Z and $\sim Z$, and pure counterfactuals are not fine-grained enough to distinguish between causal and non-causal determination relations.[13]

We have now seen both (1) that it is implausible to deny the modal interpretation of Newcomb's problem, and (2) that there are independent reasons for rejecting Gibbard and Harper's analysis of causal independence. Thus I maintain my contention that the inconsistency between these two claims should lead us to reject CI.

[12] Note also that we could remove the indexicals from G by specifying the date on which I make my choice.

[13] I do not think we need venture into the metaphysics of determination relations to accept this point. Clearly some determination relations are unproblematically non-causal, e.g. the fact that I am an unmarried male non-causally detemines that (makes true that) I am a bachelor.

III. Counterfactual Decision Theory Sanctions One-Boxing

What are the implications of rejecting CI? It shows that counterfactual decision theory in fact recommends one-boxing under the modal interpretation of Newcomb's problem. Let us see how this is so.

In the scenario, all parties admit that M is causally independent of the choice between T_1 and T_2 *in some sense*. But since we have come to deny CI, this claim does not have any immediate consequences for the probabilities of the relevant counterfactuals, so we must do some work to calculate them. To make things simple I will consider a very modally robust version of Newcomb's problem: suppose the Predictor is 100% modally reliable, so we have:

$\Box(T_1 \leftrightarrow P_1)$; and

$\Box(T_2 \leftrightarrow P_2)$.

And let us make the setup of the scenario modally robust as well:

$\Box(M \leftrightarrow P_1)$; and

$\Box(P_1 \leftrightarrow {\sim}P_2)$.

Lest these assumptions appear too extreme, note that we can restrict the necessity operator to only range over worlds that are reasonably close to ours: since we will be talking about counterfactuals with unremarkable antecedents, all we need is for these biconditionals to hold in nearby worlds. Further, similar arguments would apply for a less fully modally robust situation.[14]

Now assuming the same desirabilities as above, counterfactual decision theory holds that we calculate expected utilities as follows:

$$U(T_1) = Prob(T_1 \mathrel{\Box\!\!\rightarrow} M)(100) + prob(T_1 \mathrel{\Box\!\!\rightarrow} {\sim}M)(0)$$

$$U(T_2) = prob(T_2 \mathrel{\Box\!\!\rightarrow} M)(101) + prob(T_2 \mathrel{\Box\!\!\rightarrow} {\sim}M)(10)$$

[14] How could these suppositions be true while M is still causally independent of the prediction? First of all, as my above comments indicate, I don't think that adding the modal interpretation to the Predictor's reliability truly adds any new puzzles: the same questions arise when we are merely presented with a Predictor with high enough conditional probabilities. Further I think there are plenty of explanations of this scenario. Here is one: our world and all nearby ones are deterministic, and in these worlds the Predictor has full knowledge of the laws of nature and the facts that will affect the subject's choice at some point before the choice, and has sufficient computing power to calculate exactly what he will do. Thus in any nearby world over which our necessity operator ranges, the Predictor predicts correctly; and yet the choice itself does not cause this prediction.

What are the relevant probabilities of the counterfactuals? Consider $T_1 \Box\!\!\rightarrow M$. To determine whether it is true at our world we would need to find the smallest sphere of worlds that contains at least one T_1-world, and then see if $T_1 \supset M$ is true throughout that entire sphere. To determine the *probability* of it's truth we would generally need to make some sort of probabilistic confidence judgment about the this modal situation; however by making the scenario modally robust we avoid these difficulties. From above we have $\Box(T_1 \leftrightarrow P1)$ and $\Box(M \leftrightarrow P1)$, and from these we can obtain $\Box(T_1 \leftrightarrow M)$, and hence $\Box(T_1 \supset M)$. Thus no matter what the smallest sphere that contains a T_1-world is, we know that the material conditional $T_1 \supset M$ is true throughout it. And since a rational subject would be aware of all this (since he is aware of the setup of the problem), we have:

$$prob(T_1 \Box\!\!\rightarrow M) = 1.$$

Similar remarks apply, *mutatis mutandis*, to the other counterfactuals in the expected utility calculations, so we have:

$$prob(T_1 \Box\!\!\rightarrow \sim M) = 0;$$

$$prob(T_2 \Box\!\!\rightarrow M) = 0; \text{ and}$$

$$prob(T_2 \Box\!\!\rightarrow \sim M) = 1.$$

And so we can complete our calculations:

$$U(T_1) = (1)(100) + (0)(0) = 100$$

$$U(T_2) = (0)(101) + (1)(10) = 10.$$

Since $U(T_1) > U(T_2)$, the counterfactual decision theory developed by Gibbard and Harper endorses one-boxing, at least when the Predictor and the scenario are sufficiently modally reliable. Further, it seems plausible that this argument will go through even if this modal reliability is not 100%, so long as it is high enough to make the probabilities of the relevant counterfactuals sufficiently asymmetric to outweigh the desirability of the extra money from T_2. The "problem" (as causal decision theorists are likely to call it) is that counterfactuals only track determination relations indirectly, by means of tracking correlations across worlds. And since the modal interpretation requires that the Predictor be reliable across worlds, the mere fact that the subject's choice does not *cause* the M to be true or false is insufficient to guarantee that the truth of T_1 or T_2 can come apart from the truth of M.

IV. Conclusion

How should we react to this result? I believe that there are three salient options. First, we could deny the modal interpretation of Newcomb's problem in order to prevent the Predictor's accuracy from affecting the probabilities of the counterfactuals. However, I have already argued that this leads to some quite counterintuitive results, so I do not think that it is the most fruitful response to my argument.

Second, we could deny that Gibbard and Harper have adequately developed a version of *causal* decision theory at all. After all, we are still admitting that M is causally independent of the subject's choice *in some intuitive sense*; we are simply denying that this fact is guaranteed to have effects on the probabilities of the relevant counterfactuals. So if one's beliefs about what is rational to do in these scenarios follow this intuitive sense of causal independence, then this is a reason to reject counterfactual decision theory in favor of a different approach to representing causal relations.

Third, perhaps Gibbard and Harper were correct in their theory and wrong in their application of it: if one's intuitions about rational action are determined by the truth of the relevant counterfactuals regardless of whether they track *causal* relations, then perhaps one should admit that one-boxing is sometimes the rational choice. If this line is attractive, then it may be that some "causal" decision theorists who ascribe to counterfactual decision theory are in more agreement with evidential decision theory than was initially thought.

Though I personally am inclined toward this third option, I will not try to make a case for it here. My main conclusion is disjunctive: assuming that we accept the modal interpretation of Newcomb's problem, decision theorists must either reject counterfactual decision theory, or admit that it is sometimes rational to one-box, i.e., to act in ways that are not causally efficacious in bringing about the desired outcome.

Works Cited

Gibbard, Alan, and William L. Harper. "Counterfactuals and Two Kinds of Expected Utility." *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Ed. William L. Harper, Robert Stalnaker, and Glenn Pearce. Dordrecht: D. Reidel, 1981. 153–90.

Jeffrey, Richard. *The Logic of Decision*. 2nd ed. Chicago: Chicago UP, 1983.

Joyce, James. 2007. "Are Newcomb Problems Really Decisions?" *Synthese* 156.3 (2007): 537–62.

Lewis, David K. *Counterfactuals*. Oxford: Basil Blackwell, 1973.

——. "Prisoners' Dilemma is a Newcomb Problem." *Philosophy and Public Affairs* 8.3 (1979): 235–40.

Nozick, Robert. "Newcomb's Problem and Two Principles of Choice." *Essays in Honor of Carl G. Hempel*. Ed. Nicholas Rescher. Dordrecht: Reidel, 1969. 114–46.

Stalnaker, Robert. "A Letter to David Lewis." *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Ed. William L. Harper, Robert Stalnaker, and Glenn Pearce. Dordrecht: D. Reidel, 1981. 151–52.