Time Travel, Subjunctive Conditionals, and the Limits of Rational Choice

What can time-travel cases tell us about rational choice? In particular, what can we learn about rational choice from what we may call *retroactive choice situations*, that is, from time-travel cases in which the circumstances where an agent makes a choice depend causally on the choice the agent makes? Such cases have been thought to call into question the dominant theory of rational choice among decision theorists, namely causal decision theory (CDT). In this paper, I consider two objections that retroactive choice situations appear to raise for CDT. The first objection, which I consider in part one, is familiar from the literature. It states that CDT has false implications in many retroactive choice situations. I argue that this objection is based on highly questionable assumptions. Then in part two of the paper I raise a new objection: I argue that that CDT cannot be applied in most retroactive choice situations, since in these situations it is impossible to evaluate one's options in the manner required by causal decision theory. Thus, anyone who maintains that CDT is the correct and complete theory of rational choice must claim that, in retroactive choice situations, rational guidance is impossible.

1. First Objection: CDT has False Implications in Retroactive Choice Situations

Andy Egan has argued that causal decision theory has unacceptable implications in retroactive choice situations. After setting out his argument in section 1.1, I respond to it in section 1.2.

1.1 Egan's Time Travel Counterexample to Causal Decision Theory

According to Andy Egan, time travel cases show that causal decision theory is "fatally flawed."[1] Before setting out his objection, I must first briefly review the theory to which it is addressed. The formulation of CDT that Egan is concerned with, and that I will focus on throughout this paper, is the most standard formulation of CDT, namely the one suggested by Robert Stalnaker and developed by Allan Gibbard, William Harper, and David Lewis.[2] According to this formulation, an act is rational just in case it maximizes causal expected utility, where the causal expected utility of an act A is given by the following formula:

[1] Add reference.
[2] Add references. I will consider alternative formulations of CDT in the concluding section.

$$CEU(A) = \sum_{O} Pr(A \Rightarrow O)U(O) = Pr(A \Rightarrow O_1)U(O_1) + Pr(A \Rightarrow O_2)U(O_2) + \ldots$$

Here *CEU* stands for causal expected utility, or expected utility according to causal decision theory; *Pr* represents the agent's credences or subjective probability function; *O* ranges over the possible outcomes of the act; and *U* stands for the agent's utility function, representing how desirable the various outcomes are for the agent. The arrow ($\Rightarrow$) represents the subjunctive conditional. Hence, '$A \Rightarrow O$' should be read 'if act A were performed, then outcome O would obtain.' And so, the causal expected utility of an act is the weighted average of the utilities of the various possible outcomes, weighted by how probable it is that these outcomes would result if the action in question were performed.

Andy Egan (2007) offers the following counterexample to this theory:

> *Alexandria*: Suppose that you have a time machine, and you are convinced that time travel works in the single-timeline, no-branching way outlined by Lewis (1976). You want to use your time machine to preserve some document, thought to be lost in the fire at the library of Alexandria. One option is to send one of the grad students from your lab surreptitiously so as to remove the document from the library before the fire occurs. Another option is to send back a fleet of fire trucks, so as to prevent the fire from ever reaching the library.

Note that this case is a retroactive choice situation, since your choice situation depends causally on how you act, in the following sense. Part of your choice situation is that you know that the library was burned. Your knowing this depends on the library actually having been burned. And the library actually having been burned depends on your choosing an option that does not result in the fire's being prevented.

The Alexandria case involves three possible outcomes: you might save all the documents (if you send the fire trucks and they are successful), one of the documents (if you send the grad student and she is successful), or none of the documents, (if you choose either option and it is unsuccessful). For the sake of concreteness, suppose these three outcomes have the following utilities.

O_1 (No documents are saved):	Utility = 0
O_2 (Only one document is saved):	Utility = 1
O_3 (All documents are saved):	Utility = 20

According to Egan, it's obvious that the only rational option is to send the grad student (hereafter, GRAD), not to fire trucks (herafter, TRUCKS). He argues for this conclusion as follows:

> You know that the fire really did happen. So you know that any attempt you make to go back and prevent it will fail. It's irrational to pursue this sort of doomed plan—a plan that you already know will fail, and the failure of which you take to be worse than the expected

result of some alternative plan—and so it's irrational to try to prevent the fire.

And yet, according to Egan, CDT implies otherwise. For the causal expected utility of attempting to prevent the fire will be as follows:

$$\begin{aligned}\text{CEU}(\text{TRUCKS}) &= Pr(\text{TRUCKS} \Rightarrow O_1)U(O_1) + Pr(\text{TRUCKS} \Rightarrow O_2)U(O_2) + Pr(\text{TRUCKS} \Rightarrow O_3)U(O_3) \\ &= Pr(\text{TRUCKS} \Rightarrow O_1) \times 0 + 0 \times 1 + Pr(\text{TRUCKS} \Rightarrow O_3) \times 10 \\ &= 10 \times Pr(\text{TRUCKS} \Rightarrow O_3)\end{aligned}$$

Similarly,

$$\begin{aligned}\text{CEU}(\text{GRAD}) &= Pr(\text{GRAD} \Rightarrow O_1)U(O_1) + Pr(\text{GRAD} \Rightarrow O_2)U(O_2) + Pr(\text{GRAD} \Rightarrow O_3)U(O_3) \\ &= Pr(\text{GRAD} \Rightarrow O_1) \times 0 + Pr(\text{GRAD} \Rightarrow O_2) \times 1 + (0 \times 10) \\ &= Pr(\text{GRAD} \Rightarrow O_2)\end{aligned}$$

Thus, the causal expected utility of TRUCKS will exceed that of GRAD if and only if $Pr(\text{TRUCKS} \Rightarrow O_3)$ is more than one tenth of $Pr(\text{GRAD} \Rightarrow O_2)$. And the latter will be true so long as $Pr(\text{TRUCKS} \Rightarrow O_3) > .1$. Thus, so long as you are more than .1 confident that you'd success in preventing the fire were you to send the fire trucks, the causal expected utility of TRUCKS will exceed that of GRAD, and so CDT will recommend TRUCKS. And according to Egan, 'if you don't have a firm opinion about which course you'll actually pursue, you're likely to be confident that, if you *were* to attempt to prevent the fire [by sending the fire trucks], you would succeed. (After all, you're competent and knowledgeable, you have many willing and able accomplices, access to excellent equipment, plenty of time to plan and train, etc.)' Thus, according to Egan, if you don't have a firm opinion about which course you'll pursue, causal decision theory will recommend sending the fire trucks. Thus, according to Egan, causal decision theory will get the wrong result.

In part two, I will argue that CDT does not in fact recommend sending the fire trucks, since it is impossible to evaluate one's options in such retroactive choice situations using CDT. But regardless of whether Egan is right or wrong in thinking that causal decision theory recommends TRUCKS, he is surely right in claiming that CDT does not recommend GRAD. Hence, if GRAD is indeed the uniquely rational choice, then this would be a problem for CDT, since it would mean that CDT fails to make the correct recommendation. However, I will argue in the next section that there is strong reason to doubt the claim that GRAD is the uniquely rational choice.

1.2 How to Respond to the First Objection

The first objection to CDT can be summarized as follows.

(1) In Alexandria, you should choose GRAD.
(2) In Alexandria, CDT does not recommend GRAD.
(3) Therefore, CDT is an inadequate theory of rational choice.

I will now argue that we should not accept the conjunction of premises (1) and (2), since these premises are grounded in assumptions that are not jointly plausible.

Recall that Egan motivated premise (1) as follows: "It's irrational to pursue a plan that you already know will fail, and the failure of which you take to be worse than the expected result of some alternative plan." This claim is not entirely clear, since it's unclear how to read the locution "the expected result"—if this is taken to mean *the most probable result*, then the claim is clearly false. (After all, it can be rational to pursue plan A when the guaranteed result of plan A is worse than the most probable result of plan B, if there is some other possible result of plan B that would be disastrous). However, I think Egan is most charitably read as invoking something like the following principle.

> **Epistemic Dominance**: If one is choosing between a pair of options ϕ and ψ, and if some epistemically possible outcome of ϕ is preferable to any epistemically possible outcome of ψ, and if no epistemically possible outcome of ψ is preferable to any epistemically possible outcome of ϕ, then one is rationally required to choose ϕ over ψ.

Moreover, regardless of what Egan may have intended, Epistemic Dominance appears to be the most plausible way to support premise (1). This principle simply states the seemingly undeniable fact that, if ϕ might have a better outcome than psi, and if you know that ϕ won't have a worse outcome than ψ, then you should choose ϕ over ψ. And this principle entails premise (1). For in *Alexandria*, some epistemically possible outcome of GRAD (namely, O_2) is preferable to the only epistemically possible outcome of TRUCKS (namely, O_1), and the latter is not preferable to either of the epistemically possible outcomes of GRAD.

Turn, now, to premise (2). This premise rests on the assumption that, in *Alexandria*, what you know, in knowing that the library of Alexandria has been burned, is only that, regardless of whether you choose GRAD or TRUCKS, the fire *will* not be prevented— you do not know that, regardless of whether you choose GRAD or TRUCKS, the fire *would* not be prevented. In particular, you must not know the subjunctive conditional $(\text{TRUCKS} \Rightarrow O_1)$, for if you knew that, then the causal expected utility of TRUCKS would be zero, and so CDT would imply that you should choose GRAD. Thus, the plausibility of premise (2) appears to rest on the following assumption.

Insufficiency Principle: When someone is choosing among a set of options, and she knows, by way of backwards causation, that a certain outcome *will* not result from her choice, she does not thereby come to know that, regardless of which option she were to choose, this outcome *would* not result.

Thus, in order to undermine the first objection to CDT, it will suffice to show that Epistemic Dominance and the Insufficiency Principle are jointly implausible. And this can be shown if we consider a trio of cases, beginning with the following.

Case 1: At noon, you are holding a loaded gun, and someone offers to give you an ice cream cone if you fire the gun at your head and survive. Naturally, you decline the offer, and refrain from firing the gun at your head. Hence, at 1pm, you are alive and well, although you received no ice cream.

Clearly, in *Case 1*, at 1pm you should be glad that you refrained from firing the gun at your head. You definitely should not regret your decision. Thus, you should not think to yourself "if I knew then what I know now, then I shouldn't have refrained from firing the gun." And something similar appears to be true in the following case.

Case 2: At noon, you are holding a loaded gun, and someone offers to give you an ice cream cone if you fire the gun at your head and survive. Naturally, you decline the offer, and refrain from firing the gun at your head. At 12:59, all your memories from the past hour are erased. Hence, at 1pm, you don't remember what choice you made or whether you received the ice cream, but you know you are alive and well.

In this case, at 1pm, you can't be absolutely certain that you refrained from firing the gun. For at 1pm you should acknowledge that there's a remote chance that you fired the gun but, because of a fluke malfunction, you were unharmed. Nonetheless, you should be almost certain that you refrained from firing. And, as in Case 1, it seems clear that you should not regard your refraining from firing the gun at noon as regrettable. Thus, you should not think to yourself "if I knew then what I know now, I shouldn't have refrained from firing." After all, the only difference between Case 2 and Case 1 is that in Case 2 you can't remember how you acted, though you have strong evidence concerning how you acted. But surely your no longer remembering refraining from firing the gun can't make it the case that you should regard refraining as regrettable. Now consider.

Case 3: At noon, you are holding a loaded gun, and someone offers to give you an ice cream cone if you fire the gun at your head and survive. You have in your possession a crystal ball by which you can see one hour into the future, and you see that at 1pm you will be alive and unharmed.

Recall that in Case 2, it seems that at 1 pm you should not say to yourself "if I knew at noon what I know now, I should have fired the gun." But if the Insufficiency Principle is Correct, then

the relevant knowledge you have at 1pm in Case 2 is the same as the relevant knowledge you have at noon in Case 3. For according to the Insufficiency Principle, when, in Case 3, you look into the crystal ball and learn that you *will* not be harmed, you do not thereby learn that you *would* not be harmed regardless of which outcome you choose. Thus, what you will know in Case 3 at noon is simply that, whichever option it is that you choose at noon, you undergo no harm between noon and 1pm, which is precisely what you know at 1pm in Case 2. Consequently, if, at 1pm in Case 2, it would be false to claim that you should have fired the gun at noon if you knew then what you know at 1pm, then it must likewise be false to claim, in Case 3 (where you in fact know, at noon, what you know at 1pm in Case 2) that you should fire the gun at noon. And so we may conclude that it is not true that, at noon in Case 3, you should fire the gun.

But the Epistemic Dominance principle has the opposite implication. For in Case 3, you know that you will not be harmed in the next hour. Hence, your being harmed is not an epistemically possible outcome of your firing the, since there is no epistemically possible world in which you fire the gun and end up being harmed. And since you know that firing the gun without being harmed would result in your receiving ice cream, it follows that the only epistemically possible outcome of your firing the gun is that you are unharmed and receive ice cream. And this outcome is preferable to the only epistemically possible outcome of your refraining from firing the gun, namely, that you are unharmed and receive no ice cream. And so it follows from Epistemic Dominance that you should fire the gun.

Thus, if one accepts the Insufficiency Principle, then one should maintain that, in Case 3, it is not the case that you should fire the gun. But if one accepts Epistemic Dominance, then one must maintain that, in Case 3, it *is* the case that you should fire the gun. Therefore, we should not accept the conjunction of the Insufficiency Principle and Epistemic Dominance. And since the plausibility of Egan's objection to causal decision theory requires both these assumptions, this objection has little force.

But the causal decision theorist's troubles don't end here. For, as I will argue in what follows, retroactive choice situations present another problem for CDT, one that is much more difficult to dispel.

2. Second Objection: CDT can't be Applied to Retroactive Choice Situations

The first objection we considered to causal decision theory was that it has *false* implications in retroactive choice situations. I will now argue that CDT faces a very different difficulty, namely that it generally has *no* implications in retroactive choice situations, since it generally cannot be applied in such cases. Before making this argument, I will begin with two ground-laying sections: section 2.1 concerns a necessary condition for the applicability of CDT, and section 2.2 concerns the evaluation of the subjunctive conditionals that figure in CDT. Having laid these

foundations for the argument, in section 2.3 I will consider one particular retroactive choice situation, and I will argue that it CDT cannot be applied to it. Finally, in section 2.4, I argue that what is true in this particular case is true, in general, of retroactive choice situations, and hence that CDT, in its standard formulation, is generally inapplicable in such situations.

2.1 A Condition for the Applicability of Causal Decision Theory

Recall from section 1.1 that the causal expected utility of an act *A* is given by the following formula:

$$CEU(A) = \sum_O Pr(A \Rightarrow O)U(O) = Pr(A \Rightarrow O_1)U(O_1) + Pr(A \Rightarrow O_2)U(O_2) + \ldots$$

Thus, the causal expected utility of an act *A* is the weighted average of the utilities various possible outcomes *O*, weighted by the probabilities of the conditionals of the form $(A \Rightarrow O)$.

Now for any act *A*, the utility of this act given that $(A \Rightarrow O)$ is simply the utility of *A*. Hence, the causal expected utility of an action *A* is the weighted average of the utilities of *A* conditional on the subjunctive conditionals of the form $(A \Rightarrow O)$, weighted by the probabilities of these conditionals. But it only makes sense to calculate the expected utility of some act by averaging its utilities given various possibilities if these are *all* the possibilities. Hence, it only makes sense to calculate the expected utility of an act *A* in the manner prescribed by CDT if the subjunctive conditionals of the form $(A \Rightarrow O)$ constitute an exhaustive set of alternative possibilities. And these subjunctive conditionals will form an exhaustive set of alternative possibilities only if their probabilities sum to one. It follows that it only makes sense to apply CDT when the following condition is satisfied:

Partition Requirement: $\sum_O Pr(A \Rightarrow O) = Pr(A \Rightarrow O_1) + Pr(A \Rightarrow O_2) + \ldots = 1$

Where the Partition Requirement is violated, attempts to apply CDT yield nonsensical results.[3]

[3] This point is noted in Lewis 1973 and 1981, Joyce 1999 and Swanson (forthcoming). As one illustration of the problems that arise when this condition is violated, consider a case where an agent has two options, *A* and *B*, and there are three possible outcomes, O_1, O_2 and O_3, such that the agent prefers O_1 to O_2, and has an equally strong preference for O_2 over O_3. Suppose that act *A* is sure to result in O_2. And suppose that the Partition Requirement is violated with respect to act *B*, and, in particular, that $\Pr(B \Rightarrow O_1) + \Pr(B \Rightarrow O_2) + \Pr(B \Rightarrow O_3) = .9$. In this case, whether option *A* or *B* has the higher causal expected utility will depend on the arbitrary choice of where we set the zero point in representing the agent's utility function. Suppose we set the zero point in such a way that the utilities of O_1, O_2 and O_3 are 120, 110 and 100, respectively. In this case, the causal expected utility of *A* will 110, while the causal expected utility of *B* will be at most .9 * 120 = 108. If, however, we set the zero point so that the utilities of O_1, O_2 and O_3 are -100, -110 and -120, respectively, then the causal expected utility of *A* will be -110, while the causal expected utility of *B* will be at least .9 * (-120) = -

The Partition Requirement imposes constraints on how the outcomes, *O*, must be understood. These constraints arise from indeterministic cases, such as the following.

> *Quantum Coin*: Betty is offered a bet on an indeterministic quantum coin that has an equal chance of coming up heads or tails. If she accepts the bet and the coin comes up heads, then she wins a dollar, and if she accepts and the coin comes up tails, then she loses a dollar. But if she declines the bet, then the coin is never tossed and no money changes hands.

In this case, if Betty declines the bet, then it seems there will be no fact of the matter concerning whether he would have won or lost if he had accepted the bet. Consequently, the Partition Requirement will be violated if we define the relevant outcomes as follows.

> O_1: Betty wins a dollar
> O_2: Betty loses a dollar
> O_3: No money changes hands.

For, prior to deciding whether to accept the bet, Betty can't rule out the possibility that she will decline the bet. Hence, she can't rule out the possibility that there is no fact of the matter concerning whether she would win or lose a dollar were she to accept the bet. Hence, she cannot rule out the possibility that none of the following three conditionals is true: $(Accept \Rightarrow O_1)$; $(Accept \Rightarrow O_2)$; $(Accept \Rightarrow O_3)$. Thus, the probabilities she assigns to these three propositions must sum to less than. It follows that, if these are the relevant outcomes, then the Partition Requirement is violated.

The solution to this problem, proposed by David Lewis, is to identify outcomes, in the general case, not with maximally relevantly specific ways the world might be, but rather with *specifications of the objective chances* of such maximally relevantly specific possibilities. More precisely, we must identify outcomes with specifications of the objective chances of these possibilities once the agent's choice has been realized. Thus, in *Quantum Coin*, there will be two possible outcomes, which can be represented thus.

> O_4: After Betty's choice is realized, the objective chances are as follows:
> $Ch(O_1) = .5$; $Ch(O_2) = .5$; $Ch(O_3) = 0$
>
> O_5: After Betty's choice is realized, the objective chances are as follows:
> $Ch(O_1) = 0$; $Ch(O_2) = 0$; $Ch(O_3) = 1$

In *Quantum Coin*, Betty is fully confident that, if she were to accept the bet, then upon her doing

---

108. Thus, on the first arbitrary choice of the zero point, CDT will recommend option *B*, whereas on the second arbitrary choice of a zero point, CDT will recommend option *B*.

so there would be a .5 objective chance of O_1 and a .5 objective chance of O_2. Hence, Betty is fully confident in the following conditional: $(Accept \Rightarrow O_4)$. That is, $\Pr(Accept \Rightarrow O_4) = 1$. And she is fully confident that, if she were to decline the bet, then upon her doing so, it would be objectively certain that no money changes hands. Hence, $\Pr(Decline \Rightarrow O_5) = 1$. It follows that the Partition Requirement is satisfied. Thus, in *Quantum Coin*, this requirement can be satisfied by properly specifying the outcomes.

Let us say that an outcome is *deterministic* if it assigns a probability of one or zero to every maximally relevantly specific way the world might be, and that an outcome is *indeterministic* otherwise. Thus, O_5 is a deterministic outcome, whereas O_4 is an indeterministic outcome. Now while there may be some difference between a state of affairs obtaining and a state of affairs having an objective chance of one,[4] this difference is not relevant to our purposes. And so we may harmlessly identify a deterministic outcome with the maximally specific possibility to which it assigns a probability of one. Thus, I will identify O_5 with O_3. And similarly, in *Alexandria*, we can identify the outcome that assigns probability one to all the documents being saved with the proposition that all the documents are saved. And so we can think of the propositional outcomes that we've been considering hitherto as a limiting case of the more general class of probabilistic outcomes.

2.2 Three Principles Concerning Subjunctive Conditionals

One more issue must be discussed before presenting the second objection, namely the evaluation of subjunctive conditionals in time travel situations. This is a far from straightforward matter. There is, unfortunately, no uncontroversial theory of how counterfactuals are to be evaluated. The best-known account is the one presented by David Lewis, which involves the following *simple distance schema*:

> SDS: ($p \Rightarrow q$) is true in a world *w* just in case *w* is closer to some *p*-world than to any non-*p* world.

(Here *q* ranges over all propositions and *p* ranges over all propositions that aren't necessarily false.) Lewis conjoins this schema with an account of how to measure the distances between possible worlds,[5] thereby arriving at a general theory of subjunctive conditionals. This theory, however, is subject to numerous counterexamples. And some of these counterexamples specifically involve time travel cases.[6] It has been forcefully argued that in order to solve these problems, and in particular in order to arrive at a theory of subjunctive conditionals that applies

[4] Add references.
[5] Cite "Counterfactual Dependence and Time's Arrow."
[6] Add references.

to time travel cases and other cases involving backward causation, we need to move from the simple distance scheme to what we may call the *antecedent- relative distance schema*:[7]

> ARDS: ($p \Rightarrow q$) is true in w just in case w is closer (relative to p) to some p-world than to any non-p world.

Such antecedent-relative theories allow for the possibility that the relevant distance between two possible worlds can depend on which subjunctive conditional is being evaluated. (Hereafter I will use "closer$_p$" and "distance$_p$" to indicate the relativity of the distance to an antecedent proposition p.)

Further, there are accounts of subjunctive conditionals that make no reference whatsoever to distances between possible worlds. One example is the theory proposed by Roderick Chisholm and Nelson Goodman, according to which subjunctive conditionals can be evaluated using the *entailment schema*:

> ES: ($p \Rightarrow q$) is true in w just in case q is entailed by (p & L_w & B), where L_w is the conjunction of the laws that prevail in w, and B is the conjunction of the relevant background assumptions.

In what follows, I will adopt ARDS as a general schema for a theory of subjunctive conditionals. I can do so without much loss of generality, since each of the main theories of subjunctive conditionals is equivalent to the conjunction of ARDS and some metric for antecedent-relative distance between worlds. Thus, SDS is equivalent to ARDS in conjunction with a metric according to which the distance between two worlds relative to an antecedent p does not depend on the value of p. Similarly, ES is equivalent to ARDS conjoined with a distance metric according to which all the p-worlds that are closest$_p$ to any given world w are worlds in which (p & L_w & B).

Of course, ARDS is not a complete theory of subjunctive conditionals, but only a schema for such a theory. To arrive at a complete theory, ARDS must be combined with an account of how to compare antecedent-relative distances between worlds. I will not attempt to provide the latter, but I will defend some principles which, I will argue, ought to come out true any adequate account of antecedent-relative distance between worlds. These principles will play an important role in the arguments to come.

In stating these principles, I will make use of the notion of *causal dependence* between facts. Many authors, following Lewis, aim to explain causality in terms of subjunctive conditionals, and are therefore loath to appeal to any notion of causal dependence in their accounts of subjunctive conditionals. I needn't share these scruples, however, since my aim is not to provide a reductive account of subjunctive conditionals, but only to provide materially

[7] Add references.

adequate principles that can be used in evaluating them. Hence, I can freely appeal to the notion of causal dependence in stating these principles even if, in the final analysis, causal dependence is to be explained in terms of subjunctive conditionals.

The first principle we will need can be stated as follows.

> ***Rigidity***. For any proposition *p* and any world *w*, some *p*-worlds that match *w* with respect to all matters of particular fact that are causally independent, in *w*, of whether *p* is true are closer$_p$ to *w* than any *p*-worlds that do not so match *w*.

The argument for the Rigidity principle is simple. If proposition *q* is true and its truth is causally independent of whether *p* is true (in the sense that *p*'s being true or false has no causal bearing on *q*'s being true or false), then *q* would be true regardless of whether p is true.[8] For example, if the Hawaiian volcano Kilauea erupts, and if whether it erupts is causally independent of whether I eat Cheerios for breakfast, then Mt. Kilauea would erupt regardless of whether I eat Cheeries for breakfast. It makes no difference whether the eruption occurs before or after the would-be Cheerios consumption: so long as the eruption is causally independent of whether I eat Cheerios, it would occur regardless of whether I eat Cheeries.

But if any true proposition that is causally independent of the truth of *p* would be true regardless of whether *p* is true, then the following conditional must be true: even if *p* were false, all these causally independent propositions would be true. And so it follows from ARDS that there are some *p*-worlds where all these causally independent propositions are true that are closer$_p$ to the actual world than any *p*-worlds where it is not the case that all these causally independent propositions are true. Hence it follows that the Rigidity principle is true.

The second principle we will need concerning distances between worlds is the following:

> ***Irrelevance***. For any propositions *p* and *q* and any world *w*, if whether *q* is true in *w* depends causally on whether *p* is true in *w*, then the distance$_p$ of a given world *w*' from *w*' is not affected by whether *q* is true in *w*'.

Once again, an illustration will help to clarify this principle. Suppose that, in world *w*, Mr. Magoo is driving the wrong way on a one-way street. He notices he's approaching a Volkswagen, and so he changes lanes. Unbeknownst to Magoo, an eighteen-wheel truck is approaching in the other lane. As a result, he is hit by the truck and killed. Now consider the following subjunctive conditional: "if Magoo hadn't changed lanes, he would not have been killed." There are many facts about *w* that are relevant to the truth value of this conditional. These facts include both general laws (e.g., the laws of physics governing collisions) as well as particular facts about the situation (e.g., the velocity of Magoo's car and of the Volkswagen, the

[8] Add note: this does not apply to 'backtracking counterfactuals'. But backtracking counterfactuals are not the ones that are relevant to CDT.

condition of Magoo's airbags, etc.). The relevant facts do not, however, include the fact that Magoo was hit by an eighteen-wheeler. Thus, it would clearly be a mistake to say: "of course Mr. Magoo would have been killed if he hadn't changed lanes. After all, he was hit by an eighteen wheeler, and that'll kill someone regardless of whether they change lanes!" For Magoo was hit by an eighteen wheeler precisely because he *changed lanes*, and so this fact is not relevant to the truth value of a subjunctive conditional whose antecedent is that Magoo *did not change lanes*. More generally, where *p* is false in a world *w*, facts about *w* that are causally dependent on ~*p* are not relevant to the truth value, in *w*, of subjunctive conditionals of which *p* is the antecedent. Consequently, if we adopt ARDS as our schema for a theory of subjunctive conditionals, then we need a distance measure according to which the distance_p of a given world from *w* is not affected by the truth value in this world of propositions whose truth in *w* is causally dependent on ~*p*. In other words, we need a distance measure that satisfies the Irrelevance principle.

The third and final principle concerning subjunctive conditionals we will need for our argument is this.

> ***Legality***. Apart from similarity with respect to particular facts, what matters to the closeness of possible worlds to a given world *w* is the degree to which these worlds follow the general laws that prevail in *w*. Thus, for any proposition *p* and any worlds w_1, w_2 and w_3, if w_2 and w_3 do not differ in any relevant way in terms of how similar they are to w_1 with respect to particular matters of fact, then w_2 is closer_p than w_3 to *w* just in case the laws of w_1 are followed more closely in w_2 than in w_3.

Because this principle is a biconditional, it indicates two things. First, it indicates that miracles are not to be multiplied needlessly. As an illustration, suppose that in world *w*, Rolland rolls a normal six-sided die. And suppose the following proposition (which we may call O_6) is true: the die comes up six. And suppose that, given that the die is normal at the time when it is rolled and that no miracles occur, there are six possible outcomes of the roll: the die could come up any number between one and six. Suppose, however, that there are plenty of other metaphysically possible outcomes, but that each of these would require a miracle, that is, a violation of the laws that prevail in world *w*. Thus, there are miraculous outcomes in which the die never lands, or in which the numbers on its faces change mid-air and it comes up 42, or in which, before the die lands, it turns into a pumpkin. Nonetheless, the following subjunctive conditional is surely true: "if the die hadn't come up six, it would have come up some number between one and five." Thus, if we adopt ARDS, we must maintain that some $\neg O_6$-world where there is a non-miraculous outcome is $\text{closer}_{\neg O6}$ to the actual world than any $\neg O_6$-world where there is a miraculous outcome. It seems, therefore, that in evaluating the $\text{distance}_{\neg O6}$ between *w* and other possible worlds, we must give some weight to minimizing miracles, that is, to minimizing

violations of the laws of *w*—just as the Legality principle implies.

The second thing the Legality principle indicates is that accidental regularities, or general patterns that are not laws, are not relevant to assessing the distances between worlds. As an illustration, suppose it is an accidental regularity that all beagles born at sea belong to owners the second initial of whose name is an S. Now suppose that Fido and Rufus are two beagles belonging to J. Alfred Prufrock and Hugh Selwyn Moberly, respectively. We would not want it to come out true that if either Fido or Rufus had been born at sea, it would have been Rufus, simply because Rufus belongs to an owner whose second initial is an S. Hence, in evaluating the proximity of worlds to the actual world, we don't want to give weight to the accidental regularity that, in the actual world, all Beagles born at sea belong to owners whose second initial is an S.

So much for the three principles concerning subjunctive conditionals. In the next section, we will apply these principles to a retroactive choice situation.

2.3 An Illustration of the Inapplicability of Causal Decision Theory

Now it's time to combine the results from the last two sections. I will consider a retroactive choice situation and argue that, if we evaluate subjunctive conditionals according to the principles defended in the last section, then the Partition Requirement is violated, and so causal decision theory cannot be applied in this case.

> *The Killing Joke*: The Joker knows that Batman wants to kill him to avenge the death of his parents, and so he has devised a diabolical scheme to give Batman that opportunity. The Joker has created a time-travelling robot and programmed it to act as follows. At t_1, the robot binds Batman and the Joker with unbreakable bonds to the east and west walls of a room, respectively. At the center of the room is turret containing a laser gun that is concealed from view, though it's common knowledge that at t_1 the gun is pointing north. Within Batman's reach is a button marked "Press Me if you Dare," which he may either press or not press at t_3; neither action would require a miracle.[9] If Batman does not press the button at t_3, then at t_4 the robot releases Batman and the Joker. If, however, Batman presses the button at t_3, then at t_4 the robot activates the laser gun, so that it fires at whoever is bound to the wall at which it is pointing, and at no one else. The robot then releases whoever is left alive, and then travels back in time to t_2, whereupon it acts as follows.
>
> - If the gun is pointing east at t_3, then at t_2 the robot points the gun east.
> - If the gun is pointing west at t_3, then at t_2 the robot points the gun west.
> - If the gun is pointing neither east nor west at t_3, then at t_2 the robot points the gun

[9] Some readers might object to the supposition that neither action would require a miracle. For one might hold that, in a world governed by deterministic laws, if Batman doesn't press the button in the actual world, then his pressing the button would require a small miracle, i.e., a small violation of the laws of the actual world. I take up this objection in section 2.4, where I argue that it is based on assumptions that are false in retroactive choice situations.

either east or west, depending on the outcome of a fair coin toss.

Batman knows how the robot is programmed, and he knows that the robot is completely failsafe, and that it would take a miracle for it to fail to carry out its instructions. Not wanting to risk his own life, Batman does not press the button at t_3. Hence, at t_4, the robot releases both captives and everyone goes home.

If Batman had pressed the button at t_3, what would have been the outcome? Recall from section 1.1 that outcomes are to be understood as assignments of objective chances to maximally relevantly specific ways the world might be. For the sake of simplicity, let's assume that the maximally relevantly specific ways the world might be are as follows.

B: Batman alone is shot
J: The Joker alone is shot
N: Neither Batman alone nor the Joker alone is shot.

Thus, we'll suppose that Batman is indifferent among all the ways in which *B* might be true, and that he is also indifferent among all the ways in which *J* might be true, and that he is likewise indifferent among all the ways in which *N* might be true. It follows that, in the *Killing Joke*, each of the relevant outcomes must be a specifications of the objective chances of *B*, *J* and *N* that obtain after Batman's choice is realized. Here are two such outcomes:

O_B: After Batman's choice is realized, the objective chances are as follows: $Ch(B) = 1$; $Ch(J) = 0$; $Ch(N) = 0$

O_J: After Batman's choice is realized, the objective chances are as follows: $Ch(B) = 0$; $Ch(J) = 1$; $Ch(N) = 0$

I will now argue that if Batman had pressed the button, then the outcome would be either O_B or O_J. That is, where *p* represents the proposition that Batman presses the button, I will argue that the following conditional is true.

(1) $p \Rightarrow (O_B \vee O_J)$

I will argue, however, that it is not true that if Batman had pressed the button then the outcome would be O_B, and it is likewise not true that if Batman had pressed the button then the outcome would be O_J. That is, I will argue that neither of the following conditionals is true:

(2) $p \Rightarrow O_B$

(3) $p \Rightarrow O_J$

In outline, my argument will be this. Let *w* be the world described in the *Killing Joke*. It follows from the principles defended in the last section that the closest worlds to *w* where Batman presses the button are all and only the miracle-free worlds that match *w* with respect to all those

facts that are causally independent of whether Batman presses the button. In all of these closest worlds, $(O_B \vee O_J)$ is true. And so it follows from ARDS that (1) is true. However, in some among these closest worlds, O_B is true and O_J is false, whereas in others among these closest worlds, O_J is true and O_B is false. And so it follows from ARDS that neither (2) nor (3) is true.

Now for the detailed argument. Let's first consider proposition (1). Where *w* is the world described in the *Killing Joke*, it follows from ARDS that (1) is true just in case:

(1') There is a *p*-world where $(O_B \vee O_J)$ that is closer$_p$ to *w* than any *p*-world where $\neg(O_B \vee O_J)$.

Let's define an *initially w-matching world* as a world that matches *w* with respect to all particular matters of fact that are causally independent, in *w*, of whether *p* is true. Hence, the Rigidity principle entails that some *p*-worlds that are initially *w*-matching are closer$_p$ to *w* than any *p*-worlds that are not initially *w*-matching. Consequently, it follows from the Rigidity principle that (1') is true just in case

(1'') There is an initially *w*-matching *p*-world where $(O_B \vee O_J)$ that is closer$_p$ to *w* than any initially *w*-matching *p*-world where $\neg(O_B \vee O_J)$.

Furthermore, it follows from the Irrelevance principle that if two *p*-worlds are initially *w*-matching, then no difference between these worlds with respect to particular matters of fact is relevant to their distance$_p$ from *w*. For if two initially *w*-matching *p*-worlds differ in terms of their similarity to *w* with respect to particular matters of fact, then such differences must concern matters of fact that are causally dependent, in *w*, on whether *p* is true. And, according to the Irrelevance principle, such differences are irrelevant to the distances$_p$ of these worlds from *w*. But according to the Legality principle, if two worlds do not differ relevantly from *w* with respect to particular matters of fact, then the first of these worlds is closer$_p$ than the second to *w* just in case the laws of *w* are followed more closely in the first world than in the second. Consequently, a first initially *w*-matching *p*-world will be closer$_p$ than a second such world to *w* just in case the laws of *w* are followed more closely in the first world than in the second. Hence, (1'') will be true just in case

(1''') There is an initially *w*-matching *p*-world where $(O_B \vee O_J)$ that follows the laws of *w* more closely than any initially *w*-matching *p*-world where $\neg(O_B \vee O_J)$.

And (1''') is true, for the following reason. In *w*, while the state of the world at t_2 may be causally dependent on whether Batman presses the button at t_3 (because of the possibility of time travel from t_4 to t_2), the state of the world at t1 is not causally dependent on whether Batman presses the

button at t_3. Therefore, every initially *w*-matching *p*-world will be just like *w* with respect to the state of the world at t_1. Hence, every such world will be one in which there's a robot programmed in the manner described, and in which only a miracle (that is, only a violation of the laws of *w*) would prevent this robot from carrying out its instructions.

It follows that in every miracle-free, initially *w*-matching *p*-world, the robot carries out the instructions specified in The *Killing Joke*. Hence, in every such world, at t_2 the robot points the gun either east toward Batman or west toward the Joker. If the robot points the gun east at t_2 and Batman presses the button, then, after Batman realizes his choice, the objective chance that Batman alone is shot at t_4 is one, and so O_B is true. If, on the other hand, the robot points the gun west at t_2 and Batman presses the button, then, after Batman realizes his choice, the objective chance that the Joker alone is shot at t_4 is one, and so O_J is true. It follows that, in every miracle-free, initially *w*-matching *p*-world, the disjunction $(O_B \vee O_J)$ is true. Or, in other words, in every initially *w*-matching *p*-world in which the laws of *w* are never violated, $(O_B \vee O_J)$ is true.

Moreover, we know that there *are* miracle-free, initially *w*-matching *p*-worlds, since, *ex hypothesi*, it would not require a miracle for Batman to press the button, given the initial conditions in the *Killing Joke*. And so it follows that there are initially *w*-matching *p*-worlds where $(O_B \vee O_J)$ is true that follow the laws of *w* more closely than any initially *w*-matching *p*-worlds where $(O_B \vee O_J)$ is false. In other words, it follows that (1''') is true. And since we have seen that (1) is true just in case (1''') is true, it follows that (1) is true.

Now let's consider (2) and (3). By the same reasoning by which we concluded that (1) is true just in case (1''') is true, it also follows that (2) is true just in case

(2''') There is an initially *w*-matching *p*-world where O_B is true that follows the laws of *w* more closely than any initially *w*-matching *p*-world where O_B is false.

And it likewise follows, by the same reasoning, that (3) is true just in case

(3''') There is an initially *w*-matching *p*-world where O_J is true that follows the laws of *w* more closely than any initially *w*-matching *p*-world where O_J is false.

But (2''') and (3''') are both false. (2''') is false because there are miracle-free, initially *w*-matching *p*-worlds where O_J is true. These are worlds where, having observed that the gun is pointing west at t_3, at t_4 the robot travels back in time to t_2 and points the gun west, thereby ensuring that it fires at the Joker alone at t_4. Thus, there are initially *w*-matching *p*-worlds where O_J is true that follow the laws of *w* as closely as any such worlds where O_B is true. Hence, (2''') is false. Similarly, (3''') is false because there are miracle-free, initially *w*-matching *p*-worlds where O_B is true. These are worlds where, having observed that the gun is pointing *east* at t_3, the robot travels back to t_2 and points the gun *east*, thereby ensuring that it fires at *Batman* alone at

t_4. Thus, there are initially *w*-matching *p*-worlds where O_B is true that follow the laws of *w* as closely as any such world where O_J is true. Since (2''') and (3''') are both false, and since (2''') and (3''') if and only if (2) and (3) are true, respectively, it follows that neither (2) nor (3) is true.

To sum up, we have seen that, if in fact Batman does not press the button, then the following proposition is true.

(1) $p \Rightarrow (O_B \vee O_J)$

Moreover, we have seen that if Batman does not press the button then neither of the following propositions is true.

(2) $p \Rightarrow O_B$

(3) $p \Rightarrow O_J$

Now if (1) is true, then it follows that, apart from O_B *and* O_J, there is no relevant *O* such that $(p \Rightarrow O)$ is true. Hence, since neither (2) nor (3) is true, it follows that there is no outcome *O* such that $(p \Rightarrow O)$ is true (where O ranges over the outcomes that are relevant in the *Killing Joke*).

We are now in a position to show that the Partition Requirement is violated in the *Killing Joke*. For Batman is in a position to carry out the reasoning presented above. He should therefore recognize that, if he does not press the button, then there is no true conditional of the form $(p \Rightarrow O)$. Hence, conditional on the supposition that he does not press the button, the probabilities he assigns to these conditionals must not sum to one. Hence,

(5) $\sum_O Pr((p \Rightarrow O) | \neg p) < 1$

But according to the Partition Requirement, Batman's *unconditional* probabilities in these conditionals *must* sum to one. That is,

(6) $\sum_O Pr(p \Rightarrow O) = 1$

Now it follows from the total probability theorem[10] that we can re-express the sum of probabilities (6) as the average of the sum of these probabilities conditional on *p* and the sum of these probabilities conditional on ¬*p*, weighted by the probabilities of *p* and ¬ *p*, respectively. And so (6) is equivalent to

[10] Add reference.

(7) $$Pr(p)\sum_{O} Pr\big((p \Rightarrow O)|p\big) + Pr(\neg p)\sum_{O} Pr\big((p \Rightarrow O)|\neg p\big) = 1$$

Consider the first of the two products on the left hand side of the above equation. Since $\sum_{O} Pr\big((p \Rightarrow O)|p\big)$ cannot exceed one, it follows that this first product cannot exceed *Pr*(*p*). Therefore, (7) will be true only if the following is true.

(8) $$Pr(p) + Pr(\neg p)\sum_{O} Pr\big((p \Rightarrow O)|\neg p\big) \geq 1$$

And since *Pr*(*p*) and *Pr*(¬*p*) must sum to one, (8) will be true only if the second product on the left hand side of (8) is greater than or equal to *Pr*(¬*p*). Consequently, (8) will be true only if

(9) $$\sum_{O} Pr\big((p \Rightarrow O)|\neg p\big) \geq 1$$

But as (5) indicates, (9) is false. And so it follows that (8), (7), and (6) are likewise false. And since (6) is false, the Partition Requirement is violated. And so causal decision theory cannot be applied in the *Killing Joke*.

2.4 How the Problem Generalizes

[[[Here I generalize the argument of section 2.3, and argue that CDT will generally be inapplicable to retroactive choice situations. For all that's required to give rise to the kind of violation of the Partition Requirement we saw in section 2.3 is the following: there is some action ϕ (e.g., Batman's pressing the button) such that there is a plurality of alternative outcomes that could result from ϕ, in the absence of any miracles, and holding fixed all the facts that are causally independent of whether one ϕs. Hence, violations of the Partition Requirement will arise whenever the facts that are causally independent of how one acts are insufficient to determine a unique, nomologically possible outcome for each of one's available acts. But we should generally expect this condition to obtain in retroactive choice situations. For in such situations, the circumstances in which one acts are causally dependent on how one acts. Hence, the facts that are causally independent of how one acts are insufficient to determine the circumstances in which one acts. These facts, therefore, will generally be insufficient to determine a unique outcome for each of one's possible acts. As a result, in retroactive choice situations, the Partition Requirement will generally be violated, and so CDT will generally be inapplicable.]]]

Conclusion

[[[Here I consider the theoretical alternatives that are available in light of the preceding arguments. There are three alternatives: we can reject CDT altogether, we can reject the standard version of CDT and adopt some alternative version that dispenses with subjunctive conditionals, or we can accept the standard version of CDT and maintain that its inapplicability to retroactive choice situations is not a defect of the theory. I argue that there are serious difficulties facing the first two alternatives, and that we should opt for the third. We should deny that retroactive choice situations present soluble decision problems, and so we should deny that the correct theory of rational choice should apply to such cases.]]]