CDT and the Guaranteed Principle

Jack Spencer

Abstract

I formulate a principle of preference, which I call the Guaranteed Principle. I argue that the preferences of rational agents satisfy the Guaranteed Principle, that the preferences of agents who embody causal decision theory do not satisfy the Guaranteed Principle, and hence that causal decision theory is false.

1 Introduction

Critics of causal decision theory (CDT) have put forward various alleged counterexamples: cases in which, they claim, rationality and the recommendations of CDT diverge.[1] For the most part, proponents of CDT have been unconvinced.[2] They view the intuitions the alleged counterexamples elicit with a mixture of suspicion and opposition, and thus persist in their defence of CDT. The dispute is thus at an impasse, and one worries that unless there is some way to move beyond judgments about cases, the dispute will devolve into an unproductive clash of intuitions.

[1] See *e.g.*, Ahmed (2013; 2014a; 2014b, MS), Bostrom (2001), Egan (2007), Hare and Hedden (2016), Hunter and Richter (1978), MacAskill (2016), Oesterheld and Conitzer (MS); Price (2012), and Weirich (1985; 1988; 2004).

[2] See *e.g.* Arntzenius (2008), Cantwell (2010), Harper (1986), Joyce (2012; 2018), and Williamson (forthcoming). For relevant empirical data, see Eriksson and Rabinowicz (2013) and the studies cited therein.

My goal in this essay is move beyond the impasse. I criticize CDT, not by appeal to judgments about cases, but by explicit argument. I formulate a principle of preference, which I call the Guaranteed Principle. I argue that the preferences of rational agents satisfy the Guaranteed Principle, that the preferences of agents who embody CDT do not satisfy the Guaranteed Principle, and hence that CDT is false.

2 The Guaranteed Principle

Say that a decision *guarantees* $\$n$ if the agent knows that some particular option made available by the decision is such that she would get $\$n$ if she chose it; and say that a decision *forces* $\$n$ if the agent knows that every option made available by the decision is such that she would get $\$n$ if she chose it. If we assume that agents satisfy certain simplifying assumptions,[3] care only about money, and value dollars linearly (as I will, hereafter), then we can formulate the Guaranteed Principle as follows:

> *Guaranteed Principle*: A rational agent always strictly prefers a decision that guarantees $\$n$ to a decision that forces $\$m < \$n$.

The motivation for the Guaranteed Principle is straightforward: a rational agent should never strictly prefer fewer options. Let d_1 be a decision that forces $\$n$, and let d_2 be a decision just like d_1 except that it makes additional options available.[4] If some of the options available in d_1 are among the choiceworthy options relative to d_2, then a rational agent is indifferent between d_2 and d_1: the additional options do not improve the decision. If none of the options available in d_1 are among the most choiceworthy

[3] I assume that the agent has no self-locating uncertainty, that the agent knows that they will not suffer any information loss, that the agent knows that their utilities will not change, that the agent's utilities are bounded, and that agent's credences are conglomerable.

[4] A decision is a quadruple $\langle C, u, A, O\rangle$, where C is the credence function, u is the utility function, A is the set of options, and O is the set of possible outcomes.

options relative to d_2, then a rational agent strictly prefers d_2 to d_1: the additional options improve the decision. Either way, a rational agent weakly prefers d_2 to d_1. And a rational agent strictly prefers d_1 to some decision, d_0, which forces $\$m < \$n$. So, by transitivity, we get the Guaranteed Principle.[5]

The Guaranteed Principle does not hold of imperfect agents, nor of agents who expect to be imperfect. Take an extreme illustration. Suppose that the least choiceworthy option made available by the decision that guarantees $\$n$ is very bad indeed, and suppose that I have a lesion that makes me choose from among the least choiceworthy options when I face decisions of that sort. Then, as a way of protecting myself from my disposition to choose irrationally, I should prefer the decision that forces $\$m$ to the decision that guarantees $\$n > \$m$.

But the Guaranteed Principle, as formulated above, does not purport to hold true of imperfect agents. It's restricted to (perfectly) rational agents: the idealized agents that are the subject matter of decision theory. If an agent fully expects to choose from among the most choiceworthy options, as rational agents always do, then the agent must strictly prefer a decision that guarantees $\$n$ to a decision that forces $\$m < \$n$.

3 An Alleged Counterexample to CDT

I am going to use the Guaranteed Principle to *argue* that a particular alleged counterexample to CDT really is a counterexample. The example I will focus on is the following one, from Spencer and Wells (2019):

The Frustrater: There is an envelope and two opaque boxes, *A*

[5]The Guaranteed Principle has some obvious affinities with the causal dominance principle that features in the standard argument for two-boxing in Newcomb's problem. One-boxers who object to causal dominance reasoning might object to the Guaranteed Principle on similar, anti-causalist grounds. I defend two-boxing at length in Spencer and Wells (2019). But, in any case, my target in this essay is CDT, and proponents of CDT will not object to the Guaranteed Principle on anti-causalist grounds.

> and B. The agent has three options: she can take box A, box B, or the envelope (a_A, a_B, or a_E). The envelope contains \$40. The two boxes together contain \$100. How the money is distributed between the boxes depends on a prediction made yesterday by the Frustrater, a reliable predictor who seeks to frustrate. If the Frustrater predicted that the agent would take box A, box B contains \$100. If the Frustrater predicted that the agent would take box B, box A contains \$100. If the Frustrater predicted that the agent would take the envelope, each box contains \$50. The agent knows all of this.

There is a strong intuition that rationality requires the agent to take the envelope. CDT, however, does not recommend the envelope.

According to CDT, an agent should always choose so as to maximize U. Let C be the agent's credence function. Let $A = \{a_1, ..., a_n\}$ be the set of options.[6] Let $O = \{o_1, ..., o_m\}$ be the set of possible outcomes.[7] Let u be the agent's utility function. Let '$\Box\!\!\rightarrow$' be a nonbacktracking counterfactual conditional. The U-value of some $a \in A$, then, is:

$$U(a) = \sum_O C(a \mathrel{\Box\!\!\rightarrow} o)u(o).$$

The agent facing *The Frustrater* knows that the envelope contains \$40, so, equating dollars and units of value, $U(a_E) = 40$. The agent does not know how the money is distributed between the two boxes, but she knows that the boxes together contain \$100. Therefore, no matter how the agent divides her credence, $U(a_A) + U(a_B) = 100$.[8] Two numbers smaller than 40 cannot sum to 100, so, no matter how the agent divides her credence, a_A and/or a_B maximize U.

[6] Options are pairwise exclusive propositions the agent can make true by deciding.

[7] Possible outcomes are pairwise exclusive propositions that are, by the lights of u, unalloyed goods; *cf.* Joyce (1999).

[8] $U(a_A)+U(a_B) = (C(a_A \mathrel{\Box\!\!\rightarrow} o_0)+C(a_B \mathrel{\Box\!\!\rightarrow} o_{100})(100)+(C(a_A \mathrel{\Box\!\!\rightarrow} 0_{50})+C(a_B \mathrel{\Box\!\!\rightarrow} 0_{50})(100)+(C(a_A \mathrel{\Box\!\!\rightarrow} o_{100}) + C(a_B \mathrel{\Box\!\!\rightarrow} o_0)(100) = 100$.

Some find the intuition elicited by *The Frustrater* sufficiently compelling. They need no further argument. The case, itself, convinces them to reject CDT.

But I know—both from the literature and from personal experience—that many remain unconvinced.[9] So it's worth trying to undergird the intuition with argument.

4 Why Ain'cha Rich?

One way to try to *argue* that rationality requires taking the envelope is by appeal to a "why ain'cha rich?" argument.

Imagine agents who face *The Frustrater* repeatedly. Those who consistently take the envelope accumulate more wealth than do those who consistently take a box. A decision theory is meant to help agents get what they want in conditions of partial ignorance, and the agents we are considering care only about accumulating money. The relative poverty of box-takers as compared to envelope-takers thus might suggest that it's irrational for an agent facing *The Frustrater* to take a box.

But relative poverty does not always indicate irrationality. Consider:

> *Newcomb's Problem*: There is a transparent box and an opaque box. The agent has two options: she can take only the opaque box or both boxes ("one-box" or "two-box"). The transparent box contains $1,000. What the opaque box contains depends on a prediction made yesterday by a reliable predictor. If the predictor predicted that the agent would take both boxes, the opaque box contains $0. If the predictor predicted that the agent would take only the opaque box, the opaque box contains $1,000,000. The agent knows all of this.

[9] See *e.g.* Artnzenius (2008), Cantwell (2010), Harper (1986), Joyce (2012; 2018), and Williamson (forthcoming).

Imagine agents who face *Newcomb's Problem* repeatedly. Those who consistently one-box accumulate more wealth than do those who consistently two-box. But I think that one-boxing is irrational, nevertheless.[10]

In response to the "why ain'cha rich?" argument for one-boxing, I follow Wells (2019) and appeal to a difference in opportunity.[11] As Wells points out, "why ain'cha rich?" arguments are inferences to the best explanation. They succeed when facts about relative poverty are best explained by facts about irrationality, and they fail when facts about relative poverty are best explained otherwise. I maintain that what best explains the relative poverty of two-boxers has nothing to do with rationality and everything to do with opportunity. One-boxers have terrific opportunities; they almost always choose between $1,000,000 and $1,001,000. Two-boxers have poorer opportunities; they almost always choose between $0 and $1,000. Once we appreciate the poorer opportunities afforded to two-boxers, we should no longer be tempted to explain their relative poverty by appeal to any hypothesis concerning rationality. Irrational fools will accumulate more money than rational agents will if the opportunities afforded to the fools are better enough.

Turning back to *The Frustrater*, the pertinent question is this: what best explains the relative poverty of box-takers?

To be frank, I'm not entirely sure. There is no obvious difference in opportunity between envelope-takers and box-takers, so I suspect that the relative poverty of box-takers is best explained by the hypothesis that box-taking is irrational. But I do not know how to *argue* that no hypothesis compatible with the rationality of box-taking can explain the relative poverty of box-takers at least as well; inference to the best explanation is notoriously inconclusive.

So, although I remain bullish about the "why ain'cha rich?" argument for taking the envelope, I am going to set it aside and turn to another

[10]See Spencer and Wells (2019).

[11]For more on "why ain'cha rich?" arguments, see Ahmed (2018), Bales (2018), and Lewis (1981b).

argument, which promises to be more conclusive.

5 An Argument Against CDT

Say that an agent *embodies* a decision theory just if the agent knows that she always chooses an option recommended by the decision theory. An agent who embodies CDT knows that she always chooses a U-maximizing option. I am going to argue that rational agents do not embody CDT.

To get the argument going, consider the following elaboration of *The Frustrater*:

> *Two Rooms*: An agent must enter either Room #1 or Room #2. If she enters Room #1, she gets \$35. If she enter Room #2, she faces *The Frustrater*. Yesterday the Frustrater made a prediction about what the agent would do were she to enter Room #2. If the Frustrater predicted that the agent would take box A, box B contains \$100. If the Frustrater predicted that the agent would take box B, box A contains \$100. If the Frustrater predicted that the agent would take the envelope, each box contains \$50. The agent knows all of this.

The "decision" in Room #1 forces \$35. The decision in Room #2—namely, *The Frustrater*—guarantees \$40. The Guaranteed Principle thus entails that a rational agent strictly prefers Room #2 to Room #1.

If CDT is true, rational agents embody CDT. So we have the first premise of the argument:

P1: If CDT is true, then an agent who embodies CDT strictly prefers Room #2 to Room #1.

The second premise is a claim about the pairwise preferences of an agent who embodies CDT:

P2: An agent who embodies CDT strictly prefers Room #1 to Room #2.

To see that P2 is true, we need to run through some calculations.

Let $a_{\#1}$ and $a_{\#2}$ be the options of entering Room #1 and entering Room #2, respectively. Let o_0, o_{35}, o_{40}, o_{50}, and o_{100} be the possible outcomes of getting \$0, \$35, \$40, \$50, and \$100, respectively. We know that $U(a_{\#1}) = 35$, since Room #1 forces \$35. What $U(a_{\#2})$ is depends on how the agent divides her credence:

$$U(a_{\#2}) = C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_0)(0) + C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{35})(35) + C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{40})(40) + C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{50})(50) + C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{100})(100).$$

The agent cannot get \$35 in Room #2, and the agent knows that she would take a box were she to enter Room #2, so $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{35}) = C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{40}) = 0$. The agent's credence in the other three counterfactuals are nonzero and determined by how reliable she takes the Frustrater to be. In a more realistic case, the agent would regard the Frustrater as rather, but not perfectly, reliable. In such a case, $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_0)$ might be, say, 0.8, and $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{50})$ and $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{100})$ might be, say, 0.1. But to make the calculations simpler, suppose that the agent takes the Frustrater to be almost perfectly reliable. Then, $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{50}) \approx 0$, $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{100}) \approx 0$, and $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_0) \approx 1$. Hence:

$$U(a_{\#2}) \approx (1)(0) + (0)(35) + (0)(40) + (0)(50) + (0)(100) = 0.$$

The above calculation of $U(a_{\#2})$ relies crucially on the following fact:

(1) $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{100}) \approx 0$.

But the truth of (1) may be somewhat surprising, so let me pause here to say a bit more about it.

Let a_A, a_B, and a_E be the options made available in Room #2, and let's assume, for simplicity, that the agent thinks that box A and box B are equally likely to contain \$100. The agent knows that she would either choose box A or box B were she to enter Room #2. So,

(2) $C(a_{\#2} \boxrightarrow (a_A \vee a_B)) = 1$.

Moreover, the agent is virtually certain that one of the boxes contains \$100. So,

(3) $C(a_A \boxrightarrow o_{100}) \approx 0.5$, and

(4) $C(a_B \boxrightarrow o_{100}) \approx 0.5$.

And it might seem that (2), (3), and (4) are inconsistent with (1). If the agent thinks that both of the options she might choose were she to enter Room #2 would give her a fair chance at \$100, how can she also think that it is virtually certain that she would not get \$100 were she to enter Room #2?

But not only are (1)–(4) consistent; they're all true. The probability of a counterfactual relative to a credence function is the probability of the consequent relative to the credence function imaged on the antecedent.[12] To image a credence function on some proposition p, we take the probability assigned to any world, w, and shift it to the live p-worlds closest to w. When it comes to *Two Rooms* and *The Frustrater*, we are interested in imaging on $a_{\#2}$, a_A, or a_B; and when we image on any of these three propositions, the closeness relation hold the contents of boxes A and B fixed. When we image the agent's credence function on a_A, all of the agent's credence at worlds where box A contains \$0 is shifted to worlds where the agent get \$0, and all of the agent's credence at worlds where box A contains \$100 is shifted to worlds at which the agent get \$100; hence the truth of (3). When we image on a_B, all of the agent's credence at worlds where box B contains \$0 is shifted to worlds where the agent gets \$0, and all of the agent's credence at worlds where box B contains \$100 is shifted to worlds where the agent gets \$100; hence the truth of (4). But when we image on $a_{\#2}$, almost all of the agent's credence is shifted to worlds where the agent gets \$0; for if w is a world at which box A contains \$100, then almost all

[12] *Cf.* Lewis (1981a) and Joyce (1999).

of the closest $a_{\#2}$-worlds to w are worlds at which the agent chooses box B, and if w is a world at which box B contains \$100, then almost all of the $a_{\#2}$-worlds closest to w are worlds at which the agent chooses box A. So (1) is true: $C(a_{\#2} \mathbin{\square\!\!\rightarrow} o_{100}) \approx 0$.

If there were diachronic (conjunctive, long-arm) options, then an agent facing *Two Rooms* would have four, which we might label $a_{\#1}a_{35}$, $a_{\#2}a_A$, $a_{\#2}a_B$, and $a_{\#2}a_E$. If we assign each of these a U-value, we find that the ones that maximize U are $a_{\#2}a_A$ and/or $a_{\#2}a_B$, depending on the agent's credences. Thus, if there were such things as diachronic options, we might be able to reconcile CDT with the Guaranteed Principle. But there aren't any such things.[13] An agent deciding between Room #1 and Room #2 faces a straight choice between two (real, synchronic) options. And if the agent embodies CDT, then the agent will choose Room #1, since $U(a_{\#1}) > U(a_{\#2})$. Therefore, P2 is true.

The two premises of the argument entail the falsity of CDT. I have argued that both premises are true. So I think that we have here a sound argument against CDT.

And the underlying mistake is not hard to identify. It's not irrational for an agent who embodies CDT to strictly prefer Room #1 to Room #2—that's not where the mistake lies. After all, agents who embody CDT almost always get \$0 upon facing *The Frustrater*, and \$35 is better than \$0. The mistake lies in embodying CDT, and, specifically, in being disposed to choose so as to maximize U upon facing *The Frustrater*. An agent who knows that she will choose so as to maximize U upon facing *The Frustrater* knows that she has a poor choice-making disposition. The agent is right, then, to protect herself from that disposition by violating the Guaranteed Principle and strictly preferring Room #1 to Room #2.

But a rational agent, unlike an agent who embodies CDT, never needs to protect herself from her own choice-making dispositions. A rational agent

[13] See Hedden (2015), Joyce (1999), and Pollock (2002). For a defense of diachronic options, see McClennan (1990).

facing *Two Rooms* fully expects to take the envelope upon entering Room #2 and therefore satisfies the Guaranteed Principle, strictly preferring Room #2 to Room #1.

6 Diachronic Exploitation

Cases like *Two Rooms* reveal that the preferences of agents who embody CDT violate the Guaranteed Principle. Such cases also reveal that agents who embody CDT are diachronically exploitable. But, though I think the falsity of CDT follows from its conflict with the Guaranteed Principle, I do not think the falsity of CDT follows from the diachronic exploitability of agents who embody CDT.

Say that a sequence of options made available by a sequence of decisions *ensures* $\$n$ if the agent knows that if she took the sequence—i.e., chose each option in the sequence—she would get $\$n$; and say that an agent facing a sequence of decisions is *diachronically exploited* just if (1) there is a sequence of options available to the agent that ensures $\$n$ and (2) the agent takes a sequence that ensures $\$m < \$n$.

In *Two Rooms*, the sequence of entering Room #2 and taking the envelope ensures $40. But an agent who embodies CDT takes the "sequence" of entering Room #1, which ensures $35. So, as past critics of CDT have pointed out, agents who embody CDT are diachronically exploitable.[14]

But there are cases that convince me that diachronic exploitability and perfect rationality are compatible. One example is the following:[15]

> *Ahmed's Insurance*: At the first stage, there is an opaque box and a transparent box. The agent has two options: she can take only the opaque box or both boxes (a_1 or a_2). The transparent box contains $10. What the opaque box contains depends on a

[14]See Ahmed (2014a) and Oesterheld and Conitzer (MS).

[15]So-named because it is a variation on Ahmed's (2014a) *Newcomb Insurance*. This section also draws heavily on Ahmed (MS).

> prediction made yesterday by a reliable predictor. If the predictor predicted that the agent would take both boxes, the opaque box contains −\$50, a debt the agent must pay out of pocket. If the predictor predicted the agent would take only the opaque box, the opaque box contains \$50. At the second stage, after making the first decision but before looking into the opaque box, the agent faces a second decision. She can either bet \$75 at 1:3 that the predictor predicted correctly or bet \$25 at 3:1 that the predictor predicted incorrectly (b_1 or b_2). The agent knows all of this from the outset.

There are two possible states of the world: either the opaque box contains −\$50 or \$50 (s_{-50} or s_{50}). The agent can thus foresee the eight possible outcomes of the four possible sequences:

	s_{-50}	s_{50}
a_1b_1	−\$125	\$75
a_1b_2	\$25	\$25
a_2b_1	−\$15	−\$15
a_2b_2	−\$65	\$135

The sequence a_1b_2 ensures \$25. The sequence a_2b_1 ensures −\$15. And an agent who embodies CDT is likely to the take the sure-loss sequence, a_2b_1. The agent knows that the predictor is much more than 75% reliable whichever option is chosen. So, no matter what the agent chooses at the first stage, CDT—like any sane decision theory—recommends b_1: that the agent bet that the predictor predicted correctly at the second stage. Since the agent who embodies CDT knows that she will choose b_1 at the second stage, the U-value of taking both boxes is −15, the sure-loss value of a_2b_1, and the U-value of taking only the opaque box is:

> $U(a_1) = \sum_O C(a_1 \mathbin{\square\!\!\rightarrow} o) = C(a_1 \mathbin{\square\!\!\rightarrow} o_{75})(75) + C(a_1 \mathbin{\square\!\!\rightarrow} o_{-125})(-125) = C(s_{50})(75) + C(s_{-50})(-125)$.

Whether $U(a_2)$ or $U(a_1)$ is greater depends on how the agent divides her credence between s_{50} and s_{-50}. But suppose that the agent divides her credence equally, as she very well might. Then,

$$U(a_1) = (0.5)(75) + (0.5)(-125) = -25 < -15 = U(a_2).$$

The agent will thus take the sure-loss sequence, a_2b_1, even though a sure-gain sequence was available.

But, so far as I can tell, there is nothing irrational about taking the sure-loss sequence. To see why, it's helpful to represent the sequential decision as an intrapersonal game played between two time-slices of the agent.[16] Let C be the credence function of the first slice, and let's suppose that the credence function of the second slice comes from C via conditionalizing on the option chosen at the first stage. Let t be the proposition that the predictor predicted correctly and suppose that $C(t) = C(t|a_1) = C(t|a_2) = 0.9$. Let's also continue to suppose that $C(s_{50}) = C(s_{-50})$. We can then represent *Ahmed's Insurance* as a two-player game, using the U-values as the payoffs. In each cell, a_xb_y, of the payoff matrix below, the first coordinate is $U(a_xb_y)$ from the perspective of the first slice, i.e., $C(s_{-50})u(a_xb_ys_{-50})+C(s_{50})u(a_xb_ys_{50})$, and the second coordinate is $U(a_xb_y)$ from the perspective of the second slice, i.e., $C(s_{-50}|a_x)u(a_xb_ys_{-50}) + C(s_{50}|a_x)u(a_xb_ys_{50})$:

	b_1	b_2
a_2	$(-15, -15)$	$(35, -45)$
a_1	$(-25, 55)$	$(25, 25)$

As the payoff matrix makes clear, the game takes the form of a prisoner's dilemma.[17] Both slices want to maximize the U-value of the joint strategy

[16] Here I follow Ahmed (MS).

[17] Whether the game is a prisoner's dilemma depends on how the agent divides her credence between s_{50} and s_{-50}. But prisoner's dilemma or not: if $C(s_{-50}) > 0.45$, the only Nash equilibrium is the sure-loss sequence, a_2b_1. (If $C(s_{-50}) < 0.45$, the only Nash equilibrium is a_1b_1.)

played. The first slice maximizes the U-value of the joint strategy played by taking both boxes, no matter what the second slice does. The second slice maximizes the U-value of the joint strategy played by betting that the predictor predicted correctly, no matter what the first slice does. The two choices together lead to diachronic exploitation. But it seems to me as it will seem to many proponents of CDT: that both choices are rational.

It's worth noting that the game-theoretic perspective that helps proponents of CDT respond to the threat posed by *Ahmed's Insurance* does not help proponents of CDT respond to the threat posed by *Two Rooms*. Suppose that the agent facing *Two Rooms* thinks that box A and box B are equally likely to contain \$100. Then, if we again use the U-values of sequences as the payoffs, we get the following trivial payoff matrix:

	A	B	E
Room #1	$(35, 35)$	$(35, 35)$	$(35, 35)$
Room #2	$(50, 50)$	$(50, 50)$	$(40, 40)$

This payoff matrix does not explain why an agent who embodies CDT strictly prefers Room #2 to Room #1. In fact, it only makes the preference more puzzling; for both slices agree that the U-value of every joint strategy available in Room #2 exceeds the U-value of every joint strategy available in Room #1.

7 Conclusion

I have argued that the preferences of rational agents satisfy the Guaranteed Principle, that the preferences of agents who embody CDT do not satisfy the Guaranteed Principle, and hence that CDT is false. In so doing, I have argued that a particular alleged counterexample to CDT—namely, *The Frustrater*—really is a counterexample.

References

[1] Arif Ahmed. MS. "Sequential Choice and the Agent's Perspective."

[2] ——. 2013. "Causal Decision Theory: A counterexample." *Philosophical Review* 122: 289–306.

[3] ——. 2014a. *Evidence, Decision and Causality*. Cambridge University Press.

[4] ——. 2014b. "Dicing with Death." *Analysis* 74: 587–94.

[5] ——. 2018. "The "Why Ain'cha Rich?" Argument." In A. Ahmed (ed.) *Newcomb's Problem*, 55–73. Oxford

[6] Frank Artnzenius. 2008. "No Regrets, or: Edith Piaf revamps decision theory." *Erkenntnis* 68: 277–97.

[7] Adam Bales. 2018. "Richness and Rationality: Causal decision theory and the WAR argument." *Synthese* 195: 259–67.

[8] Nick Bostrom. 2001. "The Meta-Newcomb Problem." *Analysis* 61: 309–10.

[9] John Cantwell. 2010. "On an Alleged Counter-Example to Causal Decision Theory." *Synthese* 173: 127–52.

[10] Andy Egan. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116: 94–114.

[11] Lina Eriksson and Wlodek Rabinowicz. 2013. "The Interference Problem for the Betting Interpretation of Degrees of Belief." *Synthese* 190: 809–30.

[12] Caspar Hare and Brian Hedden. 2016. "Self-Reinforcing and Self-Frustrating Decisions." *Noûs* 50: 604–28.

[13] William Harper. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis* 24: 25–36.

[14] Brian Hedden. 2015a. *Reasons without Persons: Rationality, Identity, and Time*. Oxford University Press.

[15] ——. 2015b. "Options and Diachronic Tragedy." *Philosophy and Phenomenological Research* 90: 423–45.

[16] Daniel Hunter and Reed Richter. 1978. "Counterfactuals and Newcomb's Paradox." *Synthese* 39: 249–61.

[17] James Joyce. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press.

[18] ——. 2012. "Regret and Stability in Causal Decision Theory." *Synthese* 187: 123–45.

[19] ——. 2018. "Deliberation and Stability in Newcomb Problems and Psuedo-Newcomb Problems." In A. Ahmed (ed.), *Newcomb's Problem*. Oxford University Press.

[20] David Lewis. 1981a. "Causal Decision Theory." *Australasian Journal of Philosophy* 59: 5–30.

[21] ——. 1981b. "Why Ain'cha Rich?" *Noûs* 15: 377-80.

[22] William MacAskill. 2016. "Smokers, Psychos, and Decision-Theoretic Uncertainty." *Journal of Philosophy* 113: 425–45.

[23] Edward F. McClennen. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Oxford University Press.

[24] Caspar Oesterheld and Vincent Conitzer. MS. "Extracting Money from Causal Decision Theorists."

[25] John L. Pollock. 2002. "Rational Choice and Action Omnipotence." *Philosophical Review* 111: 1–23.

[26] Huw Price. 2012. "Causation, Chance, and the Rational Significance of Supernatural Evidence." *Philosophical Review* 121: 483–538.

[27] Jack Spencer and Ian Wells. 2019. "Why Take Both Boxes?" *Philosophy and Phenomenological Research* 99: 27-48.

[28] Paul Weirich. 1985. "Decision Instability." *Australasian Journal of Philosophy* 63: 465–72.

[29] ——. 1988. "Hierarchical Maximization of Two Kinds of Expected Utility." *Philosophy of Science* 55: 560–82.

[30] ——. 2004. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. Oxford University Press.

[31] ——. "Decisions without Sharp Probabilities." *Philosophical Scientiæ* 19: 213–25.

[32] Ian Wells. 2019. "Equal Opportunity and Newcomb's Problem." *Mind* 128:429-457.

[33] Timothy Luke Williamson. Forthcoming. "Causal Decision Theory is Safe from Psychopaths." *Erkenntnis*.