An argument against causal decision theory

Jack Spencer

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 (see e.g. Hunter and Richter 1978, Weirich 2004, Egan 2007, Ahmed 2013, 2014a, 2014b and Hare and Hedden 2016). For the most part, proponents of CDT have been unconvinced, viewing the intuitions the alleged counterexamples elicit with a mixture of suspicion and opposition.[1] The dispute is thus at an impasse, and one worries that unless there is some way to move beyond judgements about cases, the dispute will devolve into an unproductive clash of intuitions.

My goal in this paper is move beyond the impasse. I criticize CDT, not by appeal to judgements 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, 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 would yield $\$n$ if chosen; and say that a decision *forces* $\$n$ if the agent knows that every option made available by the decision would yield $\$n$ if chosen. If we assume that agents satisfy certain simplifying assumptions,[2] care only about money and value dollars linearly, 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 never strictly prefers fewer options. If d_1 is a decision that forces $\$n$, and d_2 is just like d_1 except that it makes additional options available,[3] then a rational agent weakly prefers d_2 to d_1. And a rational agent strictly prefers d_1

1 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.

2 I assume that credences are conglomerable, that utilities are bounded, that there is no self-locating uncertainty and that the agents know that their utilities will not change and that they will not suffer any information loss.

3 A decision is a quadruple $\langle C, u, A, K\rangle$, where C is the credence function, u is the utility function, A is the set of options and K is the set of dependency hypotheses.

Analysis Vol. 00 | Number 0 | July 2020 | pp. 1–10 doi: 10.1093/analys/anaa037

to some decision, d_0, which forces $\$m < \$n$. So, by transitivity, we get the Guaranteed Principle.[4]

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 a 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 does not purport to hold true of imperfect agents. It is 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 succeeds. The example I will focus on is the following one, from Spencer and Wells (2019: 34):

> (The Frustrater) There is an envelope and two opaque boxes, A and B. The agent has three options: she can take A, 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 A, then B contains \$100. If the Frustrater predicted that the agent would take B, then 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 taking the envelope. CDT, however, does not recommend the envelope.

According to CDT, an agent should always choose so as to maximize U. Let $W = \{w_1, \ldots, w_n\}$ be the set of possible worlds; let C be the agent's credence function; and let u be the agent's utility function. We then can define the

4 The Guaranteed Principle is akin to a causal dominance principle. One-boxers who object to causal dominance reasoning might object to the Guaranteed Principle on similar, anti-causalist grounds. My target in this essay is CDT, however, and proponents of CDT will not object to the Guaranteed Principle on anti-causalist grounds.

V-value of any proposition p:

$$V(p) = \sum_{W} C(w|p)u(w).$$

Note that V obeys the rule of averaging: if Z is a set of propositions that C-partitions p – in other words, if exactly one member of Z is true at every p-world to which C assigns nonzero probability – then $V(p) = \sum_Z C(z|p)V(pz)$. Let $A = \{a_1, \ldots, a_m\}$ be the set of options, and let $K = \{k_1, \ldots, k_j\}$ be the set of dependency hypotheses, where a dependency hypothesis is a maximally specific proposition about how things the agent cares about do and do not depend causally on their present choice (*Cf.* Lewis 1981: 11). The U-value of any $a \in A$, then, is:

$$U(a) = \sum_{K}\sum_{W} C(k)C(w|ak)u(w) = \sum_{K} C(k)V(ak).$$

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 boxes, but knows that the boxes together contain \$100. Therefore, no matter how the agent divides her credence, $U(a_A) + U(a_B) = 100$.[5] 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.

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 some remain unconvinced (see e.g. Joyce 2018). So it is worth trying to undergird the intuition with argument.

4. *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*:

5 There are three relevant dependency hypotheses: either A contains \$100, B contains \$100 or each box contains \$50 (k_A, k_B or k_E). $U(a_A) + U(a_B) = C(k_A)(V(a_Ak_A) + V(a_Bk_A)) + C(k_B)(V(a_Ak_B) + V(a_Bk_B)) + C(k_E)(V(a_Ak_E) + V(a_Bk_E)) = C(k_A)(100 + 0) + C(k_B)(0 + 100) + C(k_E)(50 + 50) = 100$.

(Two Rooms) An agent must enter either Room #1 or Room #2. If she enters Room #1, she gets \$35. If she enters Room #2, she faces *The Frustrater*. The agent knows all of this.[6]

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, then a rational agent embodies CDT. So we have the first premiss of the argument, which is a claim of material implication:

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

The second premiss 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 Room #2, respectively. There are three relevant dependency hypotheses: either A contains \$100, B contains \$100 or each box contains \$50 (k_A, k_B or k_E). 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:

$$(1) \quad U(a_2) = C(k_A)V(a_2k_A) + C(k_B)V(a_2k_B) + C(k_E)V(a_2k_E).$$

Let a_A, a_B and a_E be the options available in Room #2, and let us assume that each entails a_2. The agent is certain that she will choose A or B if she enters Room #2, so $\{a_A, a_B\}$ C-partitions the following propositions: a_2k_A, a_2k_B and a_2k_E. Therefore, by the rule of averaging,

$$(2) \quad V(a_2k_A) = C(a_A|a_2k_A)V(a_Ak_A) + C(a_B|a_2k_A)V(a_Bk_A);$$

$$(3) \quad V(a_2k_B) = C(a_A|a_2k_B)V(a_Ak_B) + C(a_B|a_2k_B)V(a_Bk_B); \text{ and}$$

$$(4) \quad V(a_2k_E) = C(a_A|a_2k_E)V(a_Ak_E) + C(a_B|a_2k_E)V(a_Bk_E).$$

Both $V(a_Ak_A)$ and $V(a_Bk_B)$ equal 100, since the agent gets \$100 if a_Ak_A or a_Bk_B. Both $V(a_Bk_A)$ and $V(a_Ak_B)$ equal 0, since the agent gets \$0 if a_Bk_A or a_Ak_B. And both $V(a_Ak_E)$ and $V(a_Bk_E)$ equal 50, since the agent gets \$50 if a_Ak_E or a_Bk_E. Therefore,

$$(5) \quad V(a_2k_A) = C(a_A|a_2k_A)(100) + C(a_B|a_2k_A)(0);$$

6 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 A, then B contains \$100. If the Frustrater predicted that the agent would take B, then A contains \$100. If the Frustrater predicted that the agent would take the envelope, then each box contains \$50. The agent knows that the Frustrater predicted the truth of exactly one of these three counterfactuals.

(6) $V(a_2k_B) = C(a_A|a_2k_B)(0) + C(a_B|a_2k_B)(100)$; and

(7) $V(a_2k_E) = C(a_A|a_2k_E)(50) + C(a_B|a_2k_E)(50)$.

Plugging (5)–(7) back into (1), we get:

(8) $U(a_2) = C(k_A)(C(a_A|a_2k_A)(100) + C(a_B|a_2k_A)(0))$
$+ C(k_B)(C(a_A|a_2k_B)(0) + C(a_B|a_2k_B)(100)) + C(k_E)(C(a_A|a_2k_E)(50)$
$+ C(a_B|a_2k_E)(50))$.

What these credences and conditional credences are depends on how reliable the agent takes the Frustrater to be. In a more realistic case, the agent might take the Frustrater to be rather, but not extraordinarily, reliable. But let us suppose, to make things simple, that the agent takes the Frustrater to be almost perfectly reliable. In that case, the agent is virtually certain that some box contains \$100, and virtually certain that she will take a box that contains \$0 if she enters Room #2 – in other words, $C(k_E) \approx 0$, $C(a_A|a_2k_A) \approx 0$ and $C(a_B|a_2k_B) \approx 0$. It therefore follows that:[7]

(9) $U(a_2) \approx 0$.

If there were diachronic (conjunctive, long-arm) options, then we might be able to reconcile CDT with the Guaranteed Principle. An agent facing *Two Rooms* would have four diachronic options: a_1, a_A, a_B and a_E. If we assign each of these a U-value, the ones that maximize U are a_A and/or a_B, since $U(a_1) = 35$, $U(a_E) = 40$, and $U(a_A) + U(a_B) = 100$. It is not completely clear what preferences among decisions are if there are diachronic options; but if we think of decisions as containing only synchronic options, and we think of diachronic options as having synchronic options as parts, then we could say that an agent strictly prefers decision d_i to decision d_j just if some synchronic option in d_i is a part of some diachronic option that is strictly preferred to every diachronic option that has any synchronic option in d_j as a part. An agent who embodies CDT strictly prefers a_A and/or a_B to a_1. So, if there were diachronic options, an agent who embodies CDT would, in this sense, strictly prefer Room #2 to Room #1, and the conflict between CDT and the Guaranteed Principle would be removed. But, unfortunately for CDT, there are no such things.[8] An agent deciding between Room #1 and Room #2 faces a straight choice between two (real, synchronic) options. And if the agent

7 $U(a_2) < U(a_1)$ in more realistic cases, too. For example, if $C(k_A) = C(k_B) = 0.4$, $C(a_A|a_2k_A) = C(a_B|a_2k_B) = 0.2$, and $C(a_A|a_2k_E) = C(a_B|a_2k_E) = 0.5$, then $U(a_2) = (0.4)((0.2)(100) + (0.8)(0)) + (0.4)((0.8)(0) + (0.2)(100)) + (0.2)((0.5)(50) + (0.5)(50)) = 26 < 35$.

8 See Joyce 1999, Pollock 2002 and Hedden 2015. For a defence of diachronic options, see McClennen 1990.

embodies CDT, the agent will choose Room #1, since $U(a_1) > U(a_2)$. Therefore, (P2) is true.[9]

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

And it is not hard to see where CDT goes wrong. It is not irrational for an agent who embodies CDT to strictly prefer Room #1 to Room #2 – that is 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 strong disposition to choose an empty box, and it is rational for her to protect herself from this choice-making 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 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.

5. *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.

9 CDT is sometimes formulated in terms of imaging; *cf.* Lewis 1981, Sobel 1994 and Joyce 1999. If C^a is C imaged on a, then $U(a)$ is taken to be $\sum_W C^a(w)u(w)$. Credence shifted by imaging is confined by dependency hypotheses. But a question arises: in *Two Rooms*, when we image on a_2, how do we divide the shifted credence between a_A and a_B? According to Lewis (1981), if we set time travel aside, then: for any option a and any proposition p, $C^a(p) = \sum_K C(k)C(p|ak)$. Thus, according to Lewis, credence shifted by imaging on a_2 is distributed between a_A and a_B in proportion to original credence. If Lewis's equation holds in *Two Rooms*, then my defence of (P2) carries over to imaging-based CDT straightforwardly. But, as a helpful Associate Editor pointed out, one could hold a rival view about imaging, on which the credence shifted by imaging on a_2 is divided equally between a_A and a_B, irrespective of original credence. If this rival view is accepted, my defence of (P2) does not carry over straightforwardly; for then $\sum_W \sum_K C(k)C(w|ak)u(w)$ and $\sum_W C^a(w)u(w)$ need not be equal. However, as the helpful Associate Editor also pointed out, my main contention can still be defended. If the rival view of imaging is accepted, then *Two Rooms* is unstable: $U(a_2)$ is inversely proportional to $C(a_2)$. If the agent divides their credence so that $U(a_2) = U(a_1)$, then the agent will be indifferent between Room #1 and Room #2 – thus violating the Guaranteed Principle, which requires that the agent strictly prefer Room #2 to Room #1.

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 – that is, 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 (see Ahmed 2014a, Oesterheld and Conitzer MS).

But there are cases that convince me that diachronic exploitability and perfect rationality are compatible. One example is the following:[10]

> (Ahmed's Insurance) There is a transparent box and an opaque box. The agent has two options: she can take either 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 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 repay. If the predictor predicted that the agent would take only the opaque box, the opaque box contains \$50. At the second stage, 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 relevant dependency hypotheses: either the opaque box contains \$50 or –\$50 (k_{50} or k_{-50}). The agent can thus foresee the eight possible outcomes of the four possible sequences:

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

The sequence a_1b_2 ensures \$25; the sequence a_2b_1 ensures –\$15; and an agent who embodies CDT is likely to 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 –

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

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:

$$\begin{aligned} U(a_1) &= C(k_{50})V(a_1b_1k_{50}) + C(k_{-50})V(a_1b_1k_{-50}) \\ &= C(k_{50})(75) + C(k_{-50})(-125). \end{aligned}$$

Whether $U(a_2)$ or $U(a_1)$ is greater depends on how the agent divides her credence between k_{50} and k_{-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 is helpful to represent the sequential decision as an intrapersonal game played between two time-slices of the agent.[11] Let C be the credence function of the first slice, and let us 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 us also continue to suppose that $C(k_{50}) = C(k_{-50})$. We then can 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, that is, $C(k_{50})V(a_xb_yk_{50}) + C(k_{-50})V(a_xb_yk_{-50})$, and the second coordinate is $U(a_xb_y)$ from the perspective of the second slice, that is, $C(k_{50}|a_x)V(a_xb_yk_{50}) + C(k_{-50}|a_x)V(a_xb_yk_{-50})$:

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

As the payoff matrix makes clear, the game takes the form of a prisoner's dilemma.[12] Both slices want to maximize the U-value of the joint strategy played. The first slice maximizes the U-value of the joint strategy played by

11 Here I follow Ahmed (MS).

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

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 is 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 *A* and *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 #1 to Room #2. 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.

6. *Conclusion*

I have argued that the preferences of rational agents satisfy the Guaranteed Principle, that the preferences of agents who embody CDT do not, 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.[13]

Department of Linguistics and Philosophy
Massachusetts Institute of Technology
77 Massachusetts Ave, 32d-808
Cambridge, MA 02139, USA
jackspen@mit.edu

References

Ahmed, A. (MS). Sequential choice and the agent's perspective. Unpublished manuscript.

13 For comments, questions and encouragement, I am grateful to two anonymous referees and a helpful Associate Editor; to Arif Ahmed, David Builes, Kevin Dorst, Adam Elga, Branden Fitelson, James Joyce, Sarah Moss, Agustín Rayo, Bernhard Salow, Haley Schilling and Robert Stalnaker; and to an audience at the 2020 APA Central Division.

Ahmed, A. 2013. Causal decision theory: a counterexample. *Philosophical Review* 122: 289–306.

Ahmed, A. 2014a. *Evidence, Decision and Causality*. Cambridge: Cambridge University Press.

Ahmed, A. 2014b. Dicing with death. *Analysis* 74: 587–94.

Arntzenius, F. 2008. No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis* 68: 277–97.

Cantwell, J. 2010. On an allged counter-example to causal decision theory. *Synthese* 173: 127–52.

Egan, A. 2007. Some counterexamples to causal decision theory. *Philosophical Review* 116: 94–114.

Eriksson, L. and W. Rabinowicz. 2013. The interference problem for the betting interpretation of degrees of belief. *Synthese* 190: 809–30.

Hare, C. and B. Hedden. 2016. Self-reinforcing and self-frustrating decisions. *Noûs* 50: 604–28.

Harper, W. 1986. Mixed strategies and ratifiability in causal decision theory. *Erkenntnis* 24: 25–36.

Hedden, B. 2015. Options and diachronic tragedy. *Philosophy and Phenomenological Research* 90: 423–45.

Hunter, D. and R. Richter. 1978. Counterfactuals and Newcomb's paradox. *Synthese* 39: 249–61.

Joyce, J. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Joyce, J. 2012. Regret and stability in causal decision theory. *Synthese* 187: 123–45.

Joyce, J. 2018. Deliberation and stability in Newcomb problems and pseudo-Newcomb problems. In *Newcomb's Problem*, ed. A. Ahmed, 138–59. Cambridge: Cambridge University Press.

Lewis, D. 1981. Causal decision theory. *Australasian Journal of Philosophy* 59: 5–30.

McClennen, E.F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Oxford: Oxford University Press.

Oesterheld, C. and V. Conitzer. (MS). Extracting money from causal decision theorists. Unpublished manuscript.

Pollock, J.L. 2002. Rational choice and action omnipotence. *Philosophical Review* 111: 1–23.

Sobel, J.H. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge: Cambridge University Press.

Spencer, J. and I. Wells. 2019. Why take both boxes? *Philosophy and Phenomenological Research* 99: 27–48.

Weirich, P. 2004. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. Oxford: Oxford University Press.

Williamson, T.L. forthcoming. Causal decision theory is safe from psychopaths. *Erkenntnis*.

Abstract

This paper develops an argument against causal decision theory. 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, and hence that causal decision theory is false.

Keywords: decision theory, Newcomb's problem, rational choice