An Argument Against Causal Decision Theory

Words, excluding footnotes and bibliography: 2,477

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.

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 a rational agent satisfy the Guaranteed Principle, that the preferences of an agent who embodies CDT do not satisfy the Guaranteed Principle, and hence that CDT is false.


\(^2\)See e.g. Arntzenius (2008), Cantwell (2010), Harper (1986), Joyce (2012; forthcoming), and Williamson (forthcoming). For relevant empirical data, see Eriksson and Rabinowicz (2013) and the studies cited therein.
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 a decision is such that she would get $n$ if she chose it. If we assume that agents satisfy certain simplifying assumptions, care only about money, and value dollars linearly, as I will, hereafter, then we can formulate the Guaranteed Principle as follows:

\textit{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. 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 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.

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:

---

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; see §3.
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 $m < n$.

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 (forthcoming):

*The Frustrater:* There is an envelope and two opaque boxes, $A$ 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 an agent facing *The Frustrater* 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 $A = \{a_1, \ldots, a_n\}$ be the set of options.\(^5\) Let $O = \{o_1, \ldots, o_m\}$ be the set of possible outcomes.\(^6\) Let $u$ be the agent’s utility function. Let ‘→’ be a nonbacktracking counterfactual conditional. The $U$-value of some $a \in A$, then, is:

$$U(a) = \sum_O C(a \rightarrow o)u(o).$$

---

\(^5\)Options are pairwise exclusive propositions the agent can make true by deciding; cf. Hedden (2015), Joyce (1999), Pollock (2002).

\(^6\)Possible outcomes are pairwise exclusive propositions that, by the lights of $u$, unalloyed goods; cf. Joyce (1999).
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, $U(a_A) + U(a_B) = 100$. 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 convincing. 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. 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 agent 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

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

\footnote{See e.g. Cantwell (2010; 2013), Harper (1986), Joyce (2012; forthcoming), and Williamson (forthcoming).}
opaque box contains $0. If the predictor predicted that the agent would only the opaque box, the opaque box contains $1,000,000. The agent knows all of this.

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.9

In response to the “why ain’cha rich?” argument for one-boxing, I follow Wells (forthcoming) and appeal to a difference in opportunity.10 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 deciders 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 are 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 argument, which promises to be more conclusive.

9Spencer and Wells (forthcoming) give an argument for two-boxing that I find convincing.
10Also see Ahmed (2018), Bales (2018), and Lewis (1981b).
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 enters Room #2, she faces *The Frustrater*. 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.\(^{11}\)

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

\[ P1: \text{If CDT is true, then an agent who embodies CDT strictly prefers Room #2 to Room #1.} \]

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

\[ P2: \text{An agent who embodies CDT strictly prefers Room #1 to Room #2.} \]

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 \( \text{U}(a_{#1}) = 35 \), since Room #1 forces $35. What \( \text{U}(a_{#2}) \) is depends on how the agent divides her credence:

\[
\text{U}(a_{#2}) = C(a_{#2} \rightarrow o_0) (0) + C(a_{#2} \rightarrow o_{35}) (35) + C(a_{#2} \rightarrow o_{40}) (40) + C(a_{#2} \rightarrow o_{50}) (50) + C(a_{#2} \rightarrow o_{100}) (100).
\]

\(^{11}\)We do not need the full strength of the Guaranteed Principle. Let \( d = \langle C, u, A \ O \rangle \) be some decision, and let \( \text{U}(a|d) = \sum_o C(a \rightarrow o|d) u(o) \). We can then say that \( d \) is *non-dynamic* just if, for every \( a \in A \), \( \text{U}(a) = \text{U}(a|d) \). We could restrict the Guaranteed Principle to non-dynamic decisions among decisions, since *Two Rooms* is non-dynamic.
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 
\[ C(\neg a_{#2} \rightarrow o_{35}) = C(\neg a_{#2} \rightarrow o_{40}) = 0. \]
The agent’s credence in the other three counterfactuals is 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, \( C(\neg a_{#2} \rightarrow o_5) \) might be, say, 0.8, and \( C(\neg a_{#2} \rightarrow o_{50}) \) and \( C(\neg a_{#2} \rightarrow o_{100}) \) might be, say, 0.1. But to make the calculations simpler, suppose that the agent takes the Frustrater to be almost perfectly reliable. Then, \( C(\neg a_{#2} \rightarrow o_{50}) \approx 0 \), \( C(\neg a_{#2} \rightarrow o_{100}) \approx 0 \), and \( C(\neg a_{#2} \rightarrow o_0) \approx 1 \). Hence:

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

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

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

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. \( C(a_{#2} \rightarrow (a_A \lor a_B)) = 1. \)

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

3. \( C(a_A \rightarrow o_{100}) \approx 0.5; \) and
4. \( C(a_B \rightarrow o_{100}) \approx 0.5. \)

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 are 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.\footnote{Cf. Lewis (1981a) and Joyce (1999).}
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 $B$ contains $100$ is shifted to worlds where 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 $0$; 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 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: $C(a_{\#2} \Box \rightarrow s_{100}) \approx 0$.

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.\footnote{See Hedden (2015b), Joyce (1999), and Pollock (2002). For a defense of diachronic options, see McClennan (1990).} 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.
6 Conclusion

Although Two Rooms reveals the falsity of CDT, it is not a counterexample to CDT. It is 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 will choose poorly. The agent is right, then, to protect herself from her disposition to choose irrationally in Room #2 by strictly preferring Room #1.

But a rational agent, unlike an agent who embodies CDT, never needs to protect herself from her dispositions to choose irrationally. 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.\footnote{An agent who embodies Evidential Decision Theory (EDT) knows that she always chooses a $V$-maximizing option, where $V(a) = \sum_{O} C(o|a) u(o)$, and a rational agent who embodies EDT satisfies the Guaranteed Principle. Suppose that $d_2 = \langle C, u, A, O \rangle$ guarantees $n$ and that $d_3$ forces $m$. An agent who embodies EDT, $V(d_2) = \arg\max_{A} (\sum_{O} C(o|d_2) u(o))$. And since $d_2$ guarantees $n$, $\arg\max_{A} (\sum_{O} C(o|d_2) u(o)) \geq n > m = V(d_3)$.}
References


11


