Synthese (2018) 195:259–267
https://doi.org/10.1007/s11229-016-1214-x

Richness and rationality: causal decision theory and the WAR argument

Adam Bales[1]

Received: 8 May 2016 / Accepted: 31 August 2016 / Published online: 15 September 2016

Abstract Causal decision theory (CDT) is one of our most prominent theories of rational choice and the "why ain'cha rich?" argument (WAR) is one of the most prominent objections to this theory. According to WAR, CDT is not an adequate theory of rational choice because it leads agents to make decisions that foreseeably leave them less well off than agents that decide in some other manner. Some philosophers take WAR to decisively undermine CDT. On the other hand, others (including David Lewis and Terry Horgan) take WAR to fail to resolve the debate over CDT's adequacy. In this paper, I will defend this second view: WAR does not resolve the debate at hand but instead leads to deadlock. Then, in the second half of this paper, I will show that this deadlock is not broken by a recent variant on WAR due to Caspar Hare and Brian Hedden. Not only does this result have implications for the debate over CDT's adequacy but this discussion also casts light on the broader success/rationality link.

Keywords Causal decision theory · Why ain'cha rich · Rationality · Success

1 Introduction

It is clear that practical rationality and success are closely linked: a practically rational agent will be more successful than a practically irrational agent in a wide range of circumstances. However, it is also clear that one cannot always infer from an agent's success that this agent acted rationally (for example, the fact that an agent happens to win the lottery doesn't entail that she was rational to play it). Consequently, success-based arguments for the rationality of a decision sometimes fail.

✉ Adam Bales
atb39@cam.ac.uk

[1] Trinity College, University of Cambridge, Cambridge CB2 1TQ, UK

With this in mind, in this paper I will consider a controversial success-based argument: the "why ain'cha rich?" argument (WAR) against causal decision theory (CDT), one of our most prominent theories of rational choice. On one view, this argument decisively undermines CDT (cf. Ahmed 2014, p. 130). However, on another view, WAR leads to deadlock because it isn't neutral with regard to the issue at hand and so can't resolve disagreement over CDT's adequacy.[1] Lewis (1981, p. 378), a proponent of CDT, described the situation as a "standoff". Meanwhile Horgan (1985, p. 229), an opponent of CDT, described it as a "stalemate".[2]

In the first half of this paper, I will argue in favour of the deadlock view. Then, in the second half, I will show that this deadlock is not overturned by an argument due to Hare and Hedden (2016). Not only do these results provide insight into WAR but they also cast broader light on what it would take to run a success-based argument against a theory of choice.

2 WAR and the standard debate

Start with Newcomb's problem:

> An agent is faced with a transparent box containing \$1000 and an opaque box containing either \$1,000,000 or nothing. She may either two-box (take both boxes) or one-box (take just the opaque box). However, she knows that these boxes were filled the previous day by a predictor of her behaviour with a 99% accuracy rate[3]: if this being predicted that the agent would one-box, it placed \$1,000,000 in the opaque box. Otherwise, it left this box empty.

In relation to this scenario, one prominent position holds that two-boxing is rationally required. After all, regardless of how the opaque box is filled the agent ends up \$1000 better off if she two-boxes than if she one-boxes (as she gains the \$1000 from the transparent box, as well as whatever is in the opaque box).

However, WAR targets this two-boxing position, as follows[4]:

> (P1) Agents that one-box in Newcomb's problem end up, on average, with \$990,000.
> (P2) Agents that two-box in Newcomb's problem end up, on average, with \$10,990 (that is, with $0.99 * 1000 + 0.01 * 1{,}000{,}000$).
> *Therefore*
> (P3) Agents that one-box end up, on average, far richer than agents that two-box.
> (P4) The agent facing Newcomb's problem knows this before making her decision.

[1] This is distinct from a descriptive deadlock on which WAR *doesn't* (rather than shouldn't) resolve disagreement.

[2] Horgan doesn't explicitly mention WAR but clearly thinks it leads to stalemate (see his note 4).

[3] That is, the probability that this predictor predicts that the agent will make each decision conditional on the agent actually making that very decision is 0.99.

[4] For simplicity, I speak as if agents care only (and linearly) about wealth. In note 11, I discuss a version of this argument framed in terms of the expected return to decisions (rather than the return to agents).

Therefore
(C) It is irrational for this agent to two-box.

Or, less formally, the argument goes: "Hey two-boxer, if you're so smart, why ain'cha rich?"

This argument can then be extended to undermine not just two-boxing as a decision but also CDT as a theory of choice. After all, CDT endorses two-boxing and so if this decision is not rationally permissible then CDT gives flawed guidance. To see that CDT provides this guidance, note that this theory takes a decision to be rationally required if (speaking loosely) it has a better expected causal impact than all alternative decisions (see Joyce 1999 for formal details). Further, two-boxing has the best expected causal impact in Newcomb's problem, as only this decision causes the agent to receive the $1000 in the transparent box and this is the only relevant causal effect (as the agent receives the contents of the opaque box regardless of her decision and she has no influence on the contents of this box as it was filled on the previous day).[5] Consequently, CDT labels two-boxing as rationally required and so if WAR establishes this decision's impermissibility then CDT is flawed.[6]

3 A bad argument for CDT

Lewis and Horgan thought that WAR failed to provide a neutral way into the debate about CDT's adequacy. In the next two sections, I will argue in favour of this view. I begin with a detour via an argument *for* CDT (the problem with this argument illuminates the problem with WAR). This argument starts by noting that a pre-theoretic notion of difference making is closely (perhaps analytically) connected to our concept of practical rationality. For example, we appeal to this notion when we advise a friend not to do something because it won't make a difference (or, contrariwise, when we advise her to do something because it will make a big difference). We also appeal to this notion using various other phrases as, for example, when we talk about influencing or shaping or changing things. In all of these cases, we accept that what we rationally ought to do depends on the difference that our decisions make. With the notion of difference making in mind, an argument for CDT can proceed by suggesting that difference making should be analysed in terms of causation or counterfactual dependence and noting that this gives us grounds to accept CDT (as this theory is formalised in some variants in causal terms and in other variants in counterfactual terms).

However, this is a bad argument. After all, if it is to be common ground that difference making is connected to rationality then the central dispute in decision theory just is a dispute about how we should analyse this notion. Speaking loosely, the proponent of CDT thinks this notion should be analysed in causal terms (roughly: A makes a difference to B if A causes B to be more likely to occur). Contra this, and again speaking loosely, a prominent class of CDT's opponents think that difference making should be analysed in evidential terms (roughly: A makes a difference to B if A provides evi-

[5] Newcomb's problem is stipulated to not involve backward causation.

[6] For the most part, I set aside as irrelevant for my purposes various responses to WAR (cf. Gibbard and Harper 1978, p. 153; Joyce 1999, pp. 151–154; Arntzenius 2008, pp. 289–290). However, see also note 12.

dence that B is the case).[7] To simply assume, then, in arguing for CDT, that difference making should be analysed in causal terms is effectively to assume CDT's correctness from the outset. It is for this reason that the provided argument is problematic.

4 Success and rationality

My claim will be that WAR is similarly problematic: WAR is not neutral with regards to the analysis of difference making and so undermines CDT only if we effectively assume CDT's falsity from the outset. "Deadlock", says this paper's author; "Standoff", says Lewis; "Stalemate", says Horgan.

4.1 A problem for the success/rationality link

As a first step toward this conclusion, take another diversion and imagine that each member of a group made up of black-haired people and blonde-haired people faces a decision scenario. In this scenario, each person (who knows her own hair colour) can choose to press a button labelled black or a button labelled blonde. If the person presses the button associated with her hair colour, she receives: (a) \$1,000,000 if she has black hair; and (b) \$100 if she has blonde hair. If the person presses the button not associated with her hair colour, she receives nothing.

Now, if we imagine that 99 % of both black and blonde-haired people press the button associated with their hair colour, then those who press black receive an average of \$990,000 and those that press blonde receive an average of \$99. From the comparative richness of the two groups, we might conclude that pressing black is rationally required here and hence conclude that the blonde-haired people who pressed blonde acted irrationally. Clearly, however, this conclusion is false (given that blonde-haired people receive a payout only if they press blonde). Consideration of average success has led us astray.[8] What is needed, then, is a criterion for when such considerations reflect the rationality or irrationality of the associated decisions. Fortunately, such criteria are easy enough to come by (the interesting question will be what they reveal about WAR).

4.2 Problems for WAR

Here's a clue as to how we might develop such a criterion: the unfairness of the above comparison results from the fact that, in the case at hand, black-haired people face more advantageous circumstances than blonde-haired people do (they have the

[7] Perhaps our ordinary concept of difference making is inherently causal. If so, my usage in this paper should be seen as stipulative (with this phrase being stipulated to neutrally pick out the relation between decision and world that is relevant to rational choice). The problem with the above argument will then be even more clear: too much is being read into our ordinary usage of a stipulative phrase (and circularity will occur once more if difference making is assumed to be causal in nature).

[8] Note that, unlike in Newcomb's problem, proponents of CDT and CDT's chief rival (evidential decision theory, or EDT) will agree about rationality in this case. In particular, both groups will agree that blonde-haired people ought to press the blonde button. This is what allows the case to be used to demonstrate that naive appeals to success considerations will sometimes be problematic, while remaining neutral between EDT and CDT.

opportunity to make $1,000,000 while blonde-haired people have the opportunity to make just $100). Combined with the fact that most of the black-haired people choose black (and most the blonde-haired people choose blonde), this means that the positive returns to agents that chose black partially result from these desirable circumstances (rather than from the decision). As such, it is problematic to read from this success to the rationality of the associated decision.

This suggest a necessary criterion for when such comparisons reveal decision rationality (a criterion that is neutral, insofar as it appeals to difference making but leaves open the analysis of this notion)[9]:

> FAIR: A comparison of the return to agents reflects the rationality of the decisions made by these agents (and hence is a fair comparison) only if the circumstances of all agents are the same with regards to everything except: (a) the agent's decision; and (b) those things that the agent's decision makes a difference to.[10]

FAIR is not met in the hair-colour case as the agents making each decision differ in terms of (the distribution of) their hair colour and, in this case, an agent's hair colour is neither one of her decisions nor something that her decisions make a difference to. As such, FAIR resolves the problem here.

However, given this criterion, WAR leads to deadlock. After all, whether or not WAR succeeds now depends on whether or not a comparison of all two-boxers and one-boxers satisfies FAIR. However, this criterion won't be satisfied if difference making is analysed causally. After all, the circumstances of one-boxers and two-boxers typically differs in terms of the opaque box's contents and the agent's decision does not make a difference to these contents in the relevant sense (the agent cannot causally influence these contents). Consequently, if we accept FAIR and the account of difference making associated with CDT then WAR fails. Given FAIR, then, for WAR to succeed we must reject this account of difference making and so must effectively assume CDT's falsity up front.[11] It follows that WAR is problematic in just the way that the above argument for CDT was problematic. We are left with deadlock: WAR isn't neutral with regards to CDT's truth and so doesn't resolve disagreements about CDT's adequacy.[12]

[9] Note that this criterion is necessary but not sufficient and so there are other criteria that will need to be satisfied if a comparison of agent success is to be illuminating (for example, the beliefs of the agents facing the scenario will have to be accurate in a certain sense and, perhaps following from this, the agents being compared will need to have relevantly similar beliefs). Note also that my argument will proceed by appeal to the following criteria but could be adapted to views on which success is revealed by comparing: (i) a restricted class of agents; (ii) just some of the reward received by agents; or (iii) just agents that face the same scenario.

[10] This can be weakened as: (1) it need only be that the proportion of agents facing certain circumstances is the same in each group being compared; and (2) these circumstances need only be identical in terms of those things that make a difference to the agent's payoff.

[11] If WAR is stated in terms of expected returns to decisions (rather than agent comparisons) problems still arise as either: (1) expected returns involve implicit agent comparisons anyway (if expected returns are evaluated in terms of average returns to agents); or (2) expected returns are weighted sums and WAR is again problematically circular as the debate over CDT's adequacy just is a debate about which such sum is relevant to rationality.

[12] The responses in note 6 might seem to break this deadlock in favour of CDT. However, these responses are ill suited for the purpose of breaking deadlock as they rely on premises that the opponent of CDT

Table 1 Houdini's challenge

	Predicted A	Predicted B	Predicted C
Crate A	\$1,000,000	\$0	\$0
Crate B	\$1,001,000	\$0	\$0
Crate C	\$0	\$1000	\$0

5 An argument to the contrary

However, a recent paper by Hare and Hedden (2016) claims to undermine this deadlock. In particular, Hare and Hedden agree that WAR leads to deadlock but claim to have found a similar success-based argument against CDT that avoids deadlock. However, in the remainder of this paper, I will argue that Hare and Hedden are mistaken. Deadlock is not so easily avoided.

5.1 The argument

Hare and Hedden's argument starts from a decision scenario that I will dub *Houdini's challenge* and present as follows:

> Houdini, who can perfectly predict your behaviour, fills three crates (A, B and C) with money on the basis of his prediction of your choice in this very scenario (in accordance with Table 1).[13] He then offers you the choice of any one crate.

By appeal to this scenario, Hare and Hedden then present an objection to CDT:

(P1) A self-aware, epistemically and practically-rational agent (hereafter, SEPA) will not choose C in Houdini's challenge.
(P2) If CDT is true then a SEPA will choose C in Houdini's challenge.
Therefore
(C) CDT is false.

In relation to this objection, Hare and Hedden argue for (P2) by iterated elimination of weakly-dominated strategies. That is, they first note that if CDT is true then a SEPA should be certain that she won't choose A, as this decision pays out less than choosing B if Houdini predicted A and pays out the same amount as choosing B otherwise (that is, choosing B dominates choosing A and so a SEPA will not choose A rather than

Footnote 12 continued
denies (and so don't progress the debate). Further, there are some grounds to think that these responses rely on an up-front, hidden rejection of the version of FAIR accepted by the opponent of CDT. If so then these arguments will beg the question against CDT's opponents in just the way that WAR begs the question against CDT itself. In any case, providing a detailed analysis of every possible line of argument here is beyond the scope of this paper. Instead, I will rest content with establishing a default deadlock that will be overturned only given a substantive argument that pays close attention to any hidden assumptions about FAIR.

[13] To say that Houdini can perfectly predict your behaviour is to say that your credence that Houdini predicted a certain choice, conditional on the assumption that you make that very choice, is 1.

B).[14] However, once the SEPA becomes certain that she won't choose A, she should also become certain that Houdini won't have predicted A as Houdini always predicts correctly. Then, given this new belief, the SEPA should now be certain that she won't choose B. After all, setting aside the possibility that Houdini predicted A, choosing C pays out more than choosing B if Houdini predicted B and the same amount if he predicted C (that is, choosing C now dominates choosing B and so a SEPA will not choose B rather than C). Having eliminated the possibility of choosing A or B, the SEPA is now certain that she will choose C and, it is argued, a SEPA will not surprise herself and so will choose C.[15] (P2) has been established.

In the discussion to follow, I will assume for the sake of argument that this argument for (P2) succeeds. Consequently, it is to (P1) that I turn.

5.2 WAR again/deadlock again

Hare and Hedden support (P1) by appeal to a variant on WAR.[16] This argument is simple enough: agents that choose A end up, on average, with $1,000,000 while agents that choose C end up, on average, with $0. Therefore, it might be asked of the agents that took C, "if you're so smart, why ain'cha rich?" Or, less colloquially, it might be claimed that the comparative success of agents that chose A over agents that chose C entails that taking C is not rationally permissible (and hence that a SEPA will not take C).

This argument is, of course, nearly identical to WAR: it differs only in that the same reasoning is applied to a different scenario. Consequently, it might be unclear how this argument could avoid deadlock. To answer this question, I will first outline Hare and Hedden's diagnosis of why deadlock results in the original WAR argument (using the terminology of the current paper, rather than Hare and Hedden's original terminology). Here, Hare and Hedden's starting point is largely the same as my own: FAIR.[17] In particular: (a) let C-FAIR refer to a causal version of FAIR (on which the appeal to difference making is replaced with an appeal to the causal effects of the decision); and (b) let E- FAIR refer to the evidential version of FAIR (on which the appeal to difference making is replaced with an appeal to the evidence provided by the decision). With this terminology in mind, Hare and Hedden take WAR to lead to deadlock because two-boxing is more successful than one-boxing by the comparison

[14] This reasoning relies for legitimacy on the truth of CDT, which claims that this sort of dominance reasoning is legitimate if Houdini's prediction is causally independent of the SEPA's decision (which it is). On another view, this dominance reasoning would be legitimate only if the SEPA's decision provided no evidence about Houdini's prediction (which isn't the case and so the dominance reasoning would be illegitimate).

[15] Note that this argument succeeds only if the SEPA is certain that Houdini's predictions are perfect (this weakens the argument in various ways: for example, it means that it would not withstand a move to imprecise credences).

[16] Hare and Hedden also support (P1) by appeal to intuition but I set this issue aside (my interest is in WAR).

[17] Rather than explicitly appealing to FAIR, Hare and Hedden require relevant similarity of cases and take the question of what factors make for such similarity to be what's at stake. The differences in their account are not important for my purposes.

that is appropriate according to C-FAIR[18] and yet one-boxing is more successful than two-boxing by the comparison that is appropriate according to E- FAIR.[19]

Half of the comparable result also holds in Houdini's challenge: here, if we rely on E- FAIR then choosing C is less successful than choosing A.[20] But the other half of the result does not hold: if we rely on C-FAIR then it's not the case that choosing C is more successful than choosing A (rather, these two decisions are equally successful). After all, C-FAIR requires us to hold fixed the emptiness of all three crates (when evaluating agents that choose C, as these agents are faced with empty crates and can't causally influence their contents). Such agents receive no money regardless of their decision and so all choices are equally successful.[21] So Houdini's challenge is unlike Newcomb's problem: in Houdini's challenge, it is not the case that each decision is more successful than the other according to one of the comparisons at hand (rather, C is merely equally successful to A on one of these comparisons). "No stalemate", conclude Hedden and Hare.

However, this conclusion is false. After all, contra Hedden and Hare, WAR does *not* lead to deadlock as a result of one version of FAIR leading us to label two-boxing as more successful than one-boxing and another leading us to label one-boxing as more successful than two-boxing. Rather, WAR leads to deadlock because it is not neutral with regards to the correct analysis of difference making and hence undermines CDT only if CDT's falsity is effectively assumed from the outset. To put this another way: WAR leads to deadlock because two-boxing is problematic only if we reject C-FAIR and so only if we effectively reject CDT up front (because rejecting C-FAIR is plausible only if we reject a causal account of difference making but rejecting such an account is to effectively reject CDT).

With this in mind, we can turn back to Hare and Hedden's variant on WAR. Here, deadlock arises once again. After all, as noted above, choosing C is not less successful than the alternative decisions if we accept C-FAIR and so this claim needs to be rejected upfront if the variant on WAR is to succeed. However, to reject this claim is effectively to reject CDT and so this variant on WAR succeeds only if we assume from the outset the conclusion that it sets out to establish. This alone is enough to lead to deadlock, with no need for choosing C to be *more* successful than the alternatives if C-FAIR holds.[22]

[18] This criterion requires that the compared agents face the same box contents (as the agent's decision can't causally influence these). Amongst such agents, the two-boxers gain the same reward from the opaque box as the one-boxers but also, unlike the one-boxers, gain $1000 from the transparent box and so end up richer.

[19] Because this criterion allows the opaque box's contents to vary we can run the original WAR argument.

[20] After all, this criterion allows the contents of the crates to vary and so allows Hare and Hedden's variation on WAR to proceed.

[21] Note that Hare and Hedden also compare agents who find the crates filled as agents that choose A find them filled. However, this comparison is irrelevant as CDT only recommends that agents choose C if they don't believe they're in the situation facing agents that choose A (and nothing can be concluded about subjective rationality from comparisons of agents with false beliefs of this sort). Note also that, if CDT is correct, then a SEPA *always* chooses C, despite the equal success of all decisions but this is unproblematic: there is no problem with choosing consistently given a draw.

[22] Hare and Hedden (2016, p. 626) also claim that problems arise for CDT because certain counterfactuals that hold in Newcomb's problem don't hold in Houdini's challenge (for example: the counterfactual that if the proponent of CDT had chosen as per CDT's opponent, she would have done worse than she in fact did). However, the proponent of CDT should deny that this fact truly poses a problem for CDT. In particular,

“No stalemate”, conclude Hedden and Hare but their argument here fails: “Deadlock”, says this paper’s author; “Standoff”, says Lewis; “Stalemate”, says Horgan.

6 Conclusions

At the close of the paper, here’s a careful conclusion: WAR leads to deadlock because it succeeds only if we assume, up front, the falsity of CDT. At the close of the paper, here’s a less careful conclusion. WAR is perhaps the pre-eminent argument in the philosophical literature on decision theory and so the fact that WAR leads to deadlock is not only interesting in itself but it also hints at a broader possibility: it hints at the possibility that different theorists approach decision theory from such different starting points that there will be no neutral way to settle the matter (that is, it raises the possibility that difference making is so central to practical rationality that no argument that is neutral with regards to difference making will offer a resolution of the debate). Decision-theoretic deadlock may be inescapable.

Acknowledgements Thanks to Arif Ahmed, Brian Hedden, Richard Holton, H. Orri Stefnsson and two extremely helpful reviewers for the journal.

References

Ahmed, A. (2014). *Evidence, decision and causality*. Cambridge, UK: Cambridge University Press.

Arntzenius, F. (2008). No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis*, *68*, 277–297.

Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and applications of decision theory* (pp. 125–162). Dordrecht: D. Reidel.

Hare, C., & Hedden, B. (2016). Self-reinforcing and self-frustrating decisions. *Noûs*, *50*, 604–628.

Horgan, T. (1985). Newcomb’s problem: A stalemate. In R. Campbell & L. Sowden (Eds.), *Paradoxes of rationality and cooperation: Prisoner’s dilemma and Newcomb’s problem*. Vancouver: University of British Columbia Press.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge, UK: Cambridge University Press.

Lewis, D. (1981). Why ain’cha rich? *Noûs*, *15*, 377–380.

Footnote 22 continued
the proponent of CDT should instead focus on counterfactuals related to: (a) the lack of alternatives on which she would have done better (rather than the presence of one on which she would have done worse); and (b) whether there are *any* alternatives on which this would be the case (rather than focusing on just the alternative that her opponent selects). These counterfactuals will not reveal salient differences between Newcomb’s problem and Houdini’s challenge and so there is no relevant difference between these cases from the perspective of CDT. Further, the focus on these counterfactuals is closely related to the acceptance of C-FAIR and so the counterfactuals that Hare and Hedden instead point to pose problems for CDT only if we’ve already rejected this perspective (and so an appeal to these counterfactuals cannot provide a non-question-begging argument against CDT). In any case, the argument in the body of this paper is enough to establish problems for the attack on CDT without any need for a discussion of such counterfactuals.