The Philosophical Quarterly Vol. 71, No. 4
ISSN 0031-8094

2021
doi: 10.1093/pq/pqaa086
Advance Access Publication 23rd January 2021

EXTRACTING MONEY FROM CAUSAL DECISION THEORISTS

By Caspar Oesterheld and Vincent Conitzer

Newcomb's problem has spawned a debate about which variant of expected utility maximisation (if any) should guide rational choice. In this paper, we provide a new argument against what is probably the most popular variant: causal decision theory (CDT). In particular, we provide two scenarios in which CDT voluntarily loses money. In the first, an agent faces a single choice and following CDT's recommendation yields a loss of money in expectation. The second scenario extends the first to a diachronic Dutch book against CDT.

Keywords: Newcomb's problem, causal decision theory, evidential decision theory, Dutch book arguments, ratificationism, expected utility, dynamic inconsistency, causality.

I. INTRODUCTION

In Newcomb's problem (Nozick 1969), a 'being' offers two boxes, A and B. Box A is transparent and contains $1,000. Box B is opaque and may contain either $1,000,000 or nothing. An agent is asked to choose between receiving the contents of both boxes, or of box B only. However, the being has put $1,000,000 in box B if and only if the being predicted that the agent would choose box B only. The being's predictions are uncannily accurate. What should the agent do?[1]

Causal decision theory (CDT) recommends that the agent reason as follows: I cannot causally affect the content of the boxes—whatever is in the boxes is already there. Thus, if I choose both boxes, regardless of what is in box B, I will end up with $1,000 more than if I choose only box B. Hence, I should choose both boxes.

[1] See Ahmed (2014) for a general overview of the literature on Newcomb's problem and the foundations of decision theory.

Evidential decision theory (EDT), on the other hand, recommends that the agent reason as follows: if I choose box B, then in all likelihood the being predicted that I would choose box B only, so I can expect to walk away with $1,000,000. (Even if the being is wrong some small percentage of the time, the expected value will remain at least *close* to $1,000,000.) If I choose both boxes, then I can expect to walk away with (close to) $1,000. Hence, I should choose to one-box, i.e., take only the content of box B.

One argument against CDT is that in Newcomb's problem causal decision theorists (tend to) walk away with less money than evidential decision theorists, but this argument has not proved decisive in the debate. For instance, one influential response has been that CDT makes the best out of the situation—fixing whether the money is in box B—which EDT does not (Joyce 1999, section 5.1). It would be more convincing if there were Newcomb-like scenarios in which a causal decision theorist volunteers to lose money (in expectation or with certainty).[2] Constructing such a scenario from Newcomb's problem is non-trivial. For example, in Newcomb's problem, a causal decision theorist may realise that box B will be empty. Hence, he would be unwilling to pay more than $1,000 for the opportunity to play the game.

In this paper, we provide Newcomb-like decision problems in which the causal decision theorist voluntarily loses money to another agent. We first give a single-decision scenario in which this is true only in expectation (Section II). We then extend the scenario to create a diachronic Dutch book against CDT—a two-step scenario in which the causal decision theorist is *sure* to lose money (Section III). Finally, we discuss the implications of the existence of such scenarios (Sections IV and V).

[2] Walking away with the maximum possible (expected) payoff under any circumstances is not a realistic desideratum for a decision theory: any decision theory X has a lower expected payoff than some other decision theory Y in a decision problem that rewards agents simply for using decision theory Y (cf. Skalse 2021). However, such a setup does not allow one to devise a generic scenario in which an agent voluntarily loses money, i.e., loses money in spite of having the option to walk away losing nothing.

Furthermore, scenarios with voluntary loss appear significantly more problematic for pragmatic reasons. Regardless of what you think is the right option in Newcomb's problem, you might not view Newcomb's problem as relevant ground for decision-theoretical argument because it is so unlikely that one would ever face Newcomb's problem in the real world. For instance, even if you thought that one-boxing is rational (and two-boxing is not), you might stick with CDT nonetheless because your real-world expected opportunity costs from two-boxing in Newcomb's problem are negligible. (For some discussion of this deflationary argument, see Gauthier 1989, section XI; Ahmed 2014, section 7.1.iv; Oesterheld 2019, section 1, and references therein.) However, if there is a Newcomb-like problem in which the causal decision theorist voluntarily loses money to some other agent, this generates a significant incentive to place him in such a situation.

II. EXTRACTING A PROFIT IN EXPECTATION FROM CAUSAL DECISION THEORISTS

Consider the following scenario:

> Adversarial Offer: Two boxes, B_1 and B_2, are on offer. A (risk-neutral) buyer may purchase one or none of the boxes but not both. Each of the two boxes costs \$1. Yesterday, the seller put \$3 in each box that she predicted the buyer not to acquire. Both the seller and the buyer believe the seller's prediction to be accurate with probability 0.75.[3]

For the seller to be able to predict the buyer, we here assume that no randomisation device is available to the buyer, or, more precisely, no randomisation device whose outcome the seller cannot predict with high accuracy. In Sections IV.1 and IV.4 below, we discuss variants of the problem in which the buyer is given access to an unpredictable source of randomness.

If the buyer takes either box B_i, then the expected money gained by the seller is

$$\$1 - P(\text{money in } B_i \mid \text{buyer chooses } B_i) \cdot \$3 = \$1 - 0.25 \cdot \$3 = \$0.25. \quad (1)$$

Hence, the buyer suffers an expected loss of \$0.25 (if he buys a box). The best action for the buyer therefore appears to be to not purchase either box. Indeed, this is the course of action prescribed by EDT as well as other decision theories that recommend one-boxing in Newcomb's problem (e.g. those proposed by Spohn 2012; Poellinger 2013; Levinstein and Soares 2020).

In contrast, CDT prescribes that the buyer buy one of the two boxes. Because the agent cannot causally affect yesterday's prediction, CDT prescribes to calculate the expected utility of buying box B_i as

$$P(\text{money in box } B_i) \cdot \$3 - \$1, \quad (2)$$

where $P(\text{money in box } B_i)$ is the buyer's subjective probability that the seller has put money in box B_i, *prior* to updating this belief based on his own decision. For $i = 1, 2$, let p_i be the probability that the buyer assigns to the seller having predicted him to buy B_i. Similarly, let p_0 be the probability the buyer assigns to the seller having predicted him to buy nothing. These beliefs should satisfy $p_0 + p_1 + p_2 = 1$. Because $p_0 \geq 0$, we have that $(p_0 + p_1) + (p_0 + p_2) = 2p_0 + p_1 + p_2 \geq 1$. Hence, it must be the case that $p_0 + p_1 \geq \frac{1}{2}$ or $p_0 + p_2 \geq \frac{1}{2}$ (or both). Because $P(\text{money in box } B_i) = p_0 + p_{3-i}$ for $i = 1, 2$, it is $P(\text{money in box } B_i) \geq$

[3] This decision problem resembles the widely discussed Death in Damascus scenario (introduced to the decision theory literature by Gibbard and Harper 1981, section 11) and even more closely the Frustrater case proposed by Spencer and Wells (2017), though these are not set up to result in an expected financial loss.

$\frac{1}{2}$ for at least one $i \in \{1, 2\}$. Thus, the expected utility in eq. (2) of at least one of the two possible purchases is at least $\frac{1}{2} \cdot \$3 - \$1 = \$0.50$, which is positive.

Any seller capable of predicting the causal decision theorist sufficiently well will thus have an incentive to use this scheme to exploit CDT agents. (It does not matter whether the seller subscribes to CDT or EDT.) It should be noted that even if the buyer uses CDT, his view of the deal matches the seller's as soon as the dollar is paid. That is, after observing his action, he will realise that the box he bought is empty with probability 0.75 and thus worth less than a dollar. CDT knows that it will regret its choice (see Joyce 2012; Weirich 1985, for discussions of the phenomenon of anticipated regret a.k.a. decision instability in CDT).

III. A DIACHRONIC DUTCH BOOK AGAINST CDT

Adversarial Offer results in a loss *in expectation* for the causal decision theorist. It is natural to ask whether we can use the same idea to set up a scenario in which the causal decision theorist ends up with a *sure* loss; effectively, a Dutch book.[4] Arguably, Dutch books are more convincing than scenarios with expected losses since the very meaning of 'expectations' is the subject of the debate about EDT and CDT. Of course, if the seller could predict the buyer perfectly in Adversarial Offer (instead of being right only 75% of the time), then Adversarial Offer would become a Dutch book. But can we construct a Dutch book without perfect prediction?

We have already observed that in Adversarial Offer the causal decision theorist always regrets his decision after observing its execution. This suggests the following simple approach to constructing a Dutch book. After the box is sold, the seller allows the buyer to reverse his decision for a small fee (ending up without any box and having lost only the fee). However, a CDT buyer may then anticipate eventually undoing his choice and therefore not buy a box in the first place (Ahmed 2014, section 3.2; though cf. Skyrms 1993; Rabinowicz 2000; Ahmed 2020).[5] To get our Dutch book to work, we add another choice *before* Adversarial Offer.

[4] Since our scenario will be based on Adversarial Offer, one may immediately object to the use of the term 'Dutch book' on the grounds that in a genuine Dutch book, the bookie should not know more than the agent. In Adversarial Offer, the seller has additional information about what the buyer will choose. This objection is easily taken care of, however, by using a modified version of Adversarial Offer. Instead of having the seller predict the buyer's choice, we could introduce an external predictor. The predictor fills the boxes according to the usual specification and then gives them to the seller. The seller never looks into the boxes and therefore has exactly the same information about them as the buyer.

[5] This, of course, requires that the reversal offer does not come as a surprise. Throughout, we insist that the buyer knows all the rules of the game.

ADVERSARIAL OFFER WITH OPT-OUT: It is Monday. The buyer is scheduled to face the ADVERSARIAL OFFER on Tuesday. He also knows that the seller's prediction was already made on Sunday. As a courtesy to her customer, the seller approaches the buyer on Monday. She offers to *not offer the boxes on Tuesday* if the buyer pays her \$0.20.

Note that the seller does not attempt to predict whether the buyer will pay to opt out. Also, we assume that the buyer cannot, on Monday, commit himself to a course of action to follow on Tuesday.

It seems that a rational agent should never feel compelled to accept the Monday offer. After all, doing so loses him money with certainty, whereas simply refusing both offers (on Monday and on Tuesday) guarantees that he loses no money.

CDT, however, recommends opting out on Monday, for the following reasons. A CDT buyer knows on Monday that if he does not opt out, he will buy a box on Tuesday (though he may not yet know which one). Further, he believes that whatever box he will take on Tuesday will contain \$3 with only 25% probability, thus implying an overall expected payoff of $0.25 \cdot \$3 - \$1 = -\$0.25$. This is because, on Monday, CDT treats the decision on Tuesday in the same way as it treats any other random variable in the environment. So the causal expected utility of not opting out is just what an outside observer would expect the payoff of a CDT agent facing ADVERSARIAL OFFER to be. Because this expected payoff of $-\$0.25$ is less than the certain payoff of $-\$0.20$ that can be obtained by opting out, CDT recommends opting out.

In fact, for the argument in the previous paragraph to succeed, it is only necessary that CDT is used on Tuesday; other decision theories would also recommend accepting the Monday offer, if they anticipate that the agent will use CDT on Tuesday. For instance, if the agent followed EDT on Monday and CDT on Tuesday (and is aware on Monday that he will use CDT on Tuesday), then he would still accept the Monday offer. Similarly, if the *seller* believes that the buyer will pick one of the boxes on Tuesday, then she will hope that he rejects the Monday offer. Thus, it seems that what creates the opportunity for a Dutch book is the prospect of buying a box on Tuesday (as CDT recommends), not the use of CDT on Monday.

IV. DISCUSSION

We differentiate five types of responses to these scenarios available to supporters of CDT:

(1) They could claim that these scenarios are impossible to set up, due to the requirement that the seller can predict the buyer.

(2) They could claim that even though these scenarios can be set up in principle, they are irrelevant for evaluating decision theories like CDT.
(3) They could concede that these scenarios could arise and are relevant for evaluating decision theories, but claim that CDT's recommendations in them are acceptable.
(4) They could concede that our analysis obliges them to give up on certain specific formulations of CDT, but try to modify CDT to get these scenarios right while maintaining some of its essence, in particular two-boxing and more generally the causal dominance principle.
(5) They could concede that these scenarios show that the very core of CDT (two-boxing and thus the causal dominance principle) is implausible.

We will discuss these options in turn.

IV.1 Is the scenario impossible to set up?

Surely, if one could show that a CDT agent will or can never face these scenarios—despite the seller having an obvious incentive to set them up—that would be the most convincing defence of CDT. In particular, a causal decision theorist might claim that sufficiently accurate prediction of a CDT agent is simply impossible.[6] However, not much accuracy is required, for the following reasons. The CDT agent will take one of the two boxes. Even if the seller picks the box to fill with money uniformly at random, she would therefore be right half of the time. If she can do any better than that, predicting correctly with probability $1/2 + \epsilon$, then she can extract money from the CDT agent by putting (instead of \$3) some amount between $\$2/(1 - 2\epsilon)$ and \$2 in the box predicted not to be taken. Thus, the CDT agent needs to be *completely* unpredictable in order to avoid being taken advantage of in these examples.

Most human beings are, generally speaking, at least somewhat predictable in their actions even when such predictability can be used against them. For example, in rock-paper-scissors—which structurally resembles Adversarial Offer—most people follow exploitable patterns in what moves they select (see Farber 2015, and references therein).[7] Consider such a somewhat predictable person who aims to be a causal decision theorist. It seems that he would indeed be vulnerable to the examples discussed earlier. The only defence for the supporter of CDT would seem to then be that if so, the person in question is not *truly* acting in the way that CDT describes. That is, acting according to

[6] For a general discussion of such unpredictability claims in defence of CDT, see Ahmed (2014, chapter 8).

[7] There are multiple rock-paper-scissors bots available online which attempt to predict their human opponent's future moves based on past moves (using data from other players). As of October 2020, the bot at http://www.essentially.net/rsp/ has reportedly played about 2.2 million rounds and won 61% more often than it lost.

CDT also requires being unpredictable to the seller, either by succeeding at out-thinking the seller sufficiently often, or by acting sufficiently randomly.

However, it is not always possible for the buyer to be unpredictable to the seller. For example, imagine that the buyer is a deterministic computer program whose source code is known to the seller. Then regardless of how exactly the agent works, the seller can predict the buyer's behaviour perfectly (cf. Soares and Fallenstein 2014, section 2; Cavalcanti 2010, section 5). We would thus be forced to conclude that such a program cannot possibly follow CDT, which to us is an unsatisfactory conclusion. Plausibly any other physically realised agent that chooses deterministically can at least in principle (if not with current technology) be predicted by creating or emulating an atom-by-atom copy of that agent (cf. Yudkowsky 2010, pp. 85ff.).

Nonetheless, what happens if we grant the buyer in Adversarial Offer access to a randomisation device that is unpredictable to the seller, such as a coin? First note that a causal decision theorist buyer will never *strictly* prefer choosing according to the coin toss. Indeed, he will only ever consider randomising when he is indifferent between buying B_1 and buying B_2. Thus, there are agents who abide by CDT who never randomise and instead use some tie-breaking mechanism (such as: when indifferent choose according to alphabetical order). These CDT agents are therefore susceptible to the money extraction schemes of this paper. A second issue is that in asymmetric variants of Adversarial Offer (e.g. if B_1 always contains an extra \$0.01), it is not clear that the agent should be indifferent between B_1 and B_2. In Section IV.4, we will discuss a variant of CDT (ratificationism) that explicitly requires randomisation and includes a mechanism that ensures indifference between B_1 and B_2 even in asymmetric variants of Adversarial Offer. Orthodox CDT alone, however, is insufficient to ensure that monetary loss is avoided.

We have now shown that even if we insist that a randomisation device must be available to the buyer, there are agents who abide by orthodox CDT and who are vulnerable to the money extraction scheme. Throughout the rest of this subsection, we will show that even if a randomisation device is available, *all* CDT agents are subject to money extraction schemes, including CDT agents who do randomise when indifferent and who cannot be made to have strict preferences between B_1 and B_2 by introducing asymmetries between the boxes. To such agents, the original Adversarial Offer is not a reliable money extraction scheme anymore, because the buyer might randomise and thus earn money in expectation. A natural variant of Adversarial Offer, which we will revisit in Section IV.4 below in the context of ratificationism, is to have the seller fill neither box if she predicts the buyer to randomise. After all, even if it is not predictable how the buyer's coin will come up, the decision to consult the coin itself should be just as predictable as any other (deterministic) decision. In this variant, the seller profits (in expectation) if the buyer buys a box (whether as a result of randomisation or not). However, CDT

may not recommend buying a box. This is because the buyer might believe that the seller believes that the buyer will likely randomise and thus that the seller will likely not fill any box. (Remember that a key idea in Adversarial Offer is that the buyer knows that at least one box contains money.) Under such beliefs, CDT recommends the buyer to not buy any box and thus avoid the loss of money.

Nevertheless, it is possible to construct based on Adversarial Offer money extraction schemes against CDT that work even if the agent has access to a randomisation device, favours randomisation in case of indifference, and happens to be indifferent between B_1 and B_2. The simplest is as follows. Imagine that the seller can not only predict the buyer, but can also observe (not necessarily perfectly) whether the buyer uses randomisation to decide which box to take. For example, the seller can see whether the buyer pulls out a coin and tosses it. More futuristically, we could imagine that the seller can detect the brain activity corresponding to the seller thinking 'Eeny, meeny, miny, moe, ...' or the like. Then the seller could remove all money from the boxes if she observes that the buyer selects a box at random. Apart from this causal punishment, the seller fills the boxes as usual based on a prediction about what box (if any) the buyer eventually chooses.

The key idea is that in this new scenario, the CDT-expected value of buying a box at random is negative. Choosing at random causes both boxes to (likely) be emptied. The causal expected utility of randomising is therefore about −$1 times the probability of buying a box. Since this is worse than not buying, no CDT buyer ever chooses to randomise in this variant. Despite randomisation being available, the scenario is thus equivalent to the original Adversarial Offer without randomisation. The argument from Section II therefore implies that CDT buys one of the boxes as usual and thus incurs an expected loss.

The causal decision theorist may respond that it is unrealistic to assume that the seller can observe whether the buyer determines which box to buy by randomisation. However, we see no reason why this should be considered less realistic than the ability to predict (deterministic) choice. Also, there are various other (somewhat complicated) scenarios that do not require mind-reading-type observation of whether the buyer chooses a box at random. For example, one alternative to causal punishment of randomisation is the following. Take the Adversarial Offer variant in which the seller fills no box if she predicts the buyer to randomise. Further, the seller scales up the potential content of boxes B_1 and B_2 depending on the probability that the buyer assigns to there being no money in any box. In this way, the seller could ensure that even if the buyer believes the boxes to likely be empty, the buyer will still believe the expected value of at least one of the boxes to be greater than $1 and thus buy a box whose expected value from the seller's perspective is less than $1.

IV.2 Are the scenarios irrelevant?

Even if the supporter of CDT acknowledges that these scenarios are *possible*, he might nevertheless argue that they are *irrelevant*, in the sense that the decision theory is not intended to be used for such scenarios and hence nothing that one could show about its performance in such a scenario is of significance for evaluating the theory. 'It is as if one evaluated a car by testing how it performs underwater.' There is little we can say about this response. Still, we expect it to be unattractive to most decision theorists. After all, our scenarios (in particular Adversarial Offer) resemble Newcomb's problem—the problem that is supposed to positively distinguish CDT. Further, if our scenarios were out of CDT's scope, then we (and presumably most other decision theorists) would still be interested in identifying a decision theory that *does* make good recommendations for predictable agents (such as artificial intelligent agents whose behaviour is determined by a computer program) facing a wide range of scenarios including the ones given in this paper.

IV.3 Can CDT's recommendations be defended?

If our scenarios are within the scope of CDT, then the supporter of CDT has to contend with the fact that one can extract expected money from, and even Dutch-book, CDT agents in them. (See footnote 4 on the use of the term Dutch book to describe Adversarial Offer with Opt-Out.) But he might question the significance of Dutch book arguments and other money extraction schemes, either in general or in this particular context. For some general discussion of whether (diachronic) Dutch books are conclusive decision-theoretic arguments, see Vineberg (2016) or Hájek (2009). Note, though, that some of the most influential arguments in favour of expected utility maximisation—of which CDT is a refinement—are Dutch books. Of course, one might adopt expected utility maximisation for reasons other than Dutch books. But it would seem odd to follow Dutch book arguments to expected utility maximisation but no further.

Instead of rehashing some of the more generic reasons for and against the persuasiveness of Dutch books and loss of money in expectation, we here discuss a response that is specific to CDT and Adversarial Offer with Opt-Out.[8] A causal decision theorist may argue that it is not generally fair to expect any kind of coherence from CDT's recommendations when multiple decisions are to be made across time, due to the different perspectives that the decision maker adopts (and, arguably, has to adopt) at different points in time. Consider Newcomb's problem. Let t_0 be the time at which the predictor

[8] For a discussion of similar arguments about other diachronic Dutch books, see Rabinowicz (2008).

observes the agent (perhaps using functional magnetic resonance imaging (fMRI) or the like) in order to make a prediction. Then, before t_0, CDT recommends committing—and if needed paying money to commit—to one-boxing (cf. Barnes 1997; Joyce 1999, pp. 153f.; Meacham 2010). After t_0, CDT recommends two-boxing. However, most decision theorists do not consider this to be a compelling argument against CDT. The causal decision theorist can easily justify the difference in the decision made by the fact that, before t_0, the commitment decision has a causal effect on what is in the boxes, and after t_0, it does not.

It would be hypocritical for an evidential decision theorist to disagree, since EDT is dynamically inconsistent in analogous ways. For instance, consider a version of Newcomb's problem in which both boxes are transparent. Let t_0' be the time at which the EDT agent sees the content of both boxes. Then before t_0', EDT recommends committing—and if needed paying money to commit—to one-boxing. After t_0', EDT recommends two-boxing.[9] The evidential decision theorist can easily justify this along similar lines: before t_0', her commitment is evidence about what is in the boxes, and after t_0' it no longer is.

Thus, at least some types of dynamic inconsistency do not constitute strong arguments against a decision theory. However, in our opinion, the dynamic inconsistency displayed by CDT in Adversarial Offer with Opt-Out is much more problematic. For one, it leads to a Dutch book. Often, the main argument that is given for why a particular inconsistency is problematic is precisely that it allows for a Dutch book (Hájek 2009, section 4). Conversely, defences of dynamic inconsistencies (see Ahmed 2014, section 3.2, for an example in a Newcomb-like scenario) often focus on arguing that they do *not* allow for Dutch Books.

Further, it seems that some of the reasons for (or defences of) dynamic inconsistency in the above decision problems do not apply to CDT's dynamic inconsistency in Adversarial Offer with Opt-Out. For CDT in Newcomb's problem, there is a particular event at time t_0 that splits the decision perspectives: the loss of causal control at t_0 over the content of box B. Similarly, for EDT in the Newcomb's problem with transparent boxes, that event is the loss of *evidential* control (cf. Almond 2010, section 4.5) at t_0' over the content of box B. It is thus easy to argue for defenders of the respective theories that the perspectives from before and after t_0 or t_0' *should* diverge (Ahmed and Price 2012, pp. 23f, section 4). In sharp contrast, Adversarial Offer with Opt-Out lacks any such event between the decision points. The difference in

[9] To our knowledge, Gibbard and Harper (1981, section 10) first proposed this transparent version of Newcomb's problem (for further discussion, see Gauthier 1989; Drescher 2006, section 6.2; Arntzenius 2008, section 7; Meacham 2010, section 3.2.2). Parfit's (1984) hitchhiker (Barnes 1997), XOR Blackmail (Levinstein and Soares 2020, section 2) and Yankees vs. Red Sox (Arntzenius 2008; Ahmed and Price 2012, pp. 23f) similarly expose dynamic inconsistencies in EDT. Conitzer (2015) gives a somewhat different type of scenario—based on the Sleeping Beauty problem—in which EDT is dynamically inconsistent.

perspectives for CDT appears to be purely a result of CDT viewing its current choice differently than it views past and future decisions.

All that being said, caution should be taken when evaluating a decision theory based on scenarios with multiple decisions across time. In general, more research on what conclusions can be drawn from such scenarios is needed (cf. Steele and Stefánsson 2016, section 6). Nevertheless, we do not see any clear path by which such research would justify CDT's recommendations in Adversarial Offer with Opt-Out. In any case, even if one is at this point unwilling to consider scenarios with multiple decision points at all for the purpose of evaluating decision theories, one would still have to contend with the simpler Adversarial Offer scenario, in which there is only one decision point.

IV.4 Modifying CDT to avoid money extraction schemes

If a straightforward interpretation of CDT cannot be defended against our scenarios, one may look to modify it to avoid expected or sure loss while preserving some of CDT's core tenets. In particular, in response to other alleged counterexamples, some authors have tried to modify CDT while maintaining the causal dominance (Joyce 1999, section 5.1) a.k.a. sure thing (Gibbard and Harper 1981, section 7) principle (though see Ahmed 2012, for an argument against the motivation behind some of these approaches).

Ratificationism. For example, one may turn to the concept of ratifiability. In Newcomb-like scenarios such as those under discussion here, for any choice a, we can consider the beliefs about what is in the boxes that would result from knowing that one will choose a. Then, a choice a is ratifiable if it is an optimal choice—as judged by CDT—under those beliefs. For example, in Newcomb's problem only two-boxing is ratifiable, precisely because it is causally dominant. For an overview of ratification and its relation to CDT, see Weirich (2016, section 3.6). Unfortunately, this concept is of no help in Adversarial Offer, because none of the three options (buying B_1, buying B_2 or declining) is ratifiable. For instance, under the beliefs that would result from knowing that one will take box B_i, it would be better to buy the other box B_{3-i}.

The ratificationist may respond by claiming (as in Section IV.1) that unpredictable randomisation should always be possible. If that were true, then the only ratifiable option would be to take each box with probability 50%, thus gaining money in expectation. But again, we would like to have a decision theory that works in a broad variety of scenarios, including ones where the agent expects to be somewhat predictable. Furthermore, even if a source of true randomness (i.e. randomness unpredictable to the seller) is in fact available, this does not settle the issue. For example, consider (again) a variant of Adversarial Offer in which the seller refrains from putting money in any box if she

predicts the buyer to make different choices depending on the randomisation device.[10] In this variant of the problem, again no option is ratifiable: under the beliefs that would result from knowing that the buyer will choose at random (and therefore choose a box with some positive probability), the buyer would rather not pick any box. To circumvent this example, the ratificationist could argue that the decision maker should be able to randomise in such a way that *whether* he is randomising is unpredictable. However, at this point, one might just as well assert the impossibility or irrelevance of Newcomb-type scenarios altogether, which we have addressed in Sections IV.1 and IV.2.

Policy choice. A different strategy for modifying CDT to avoid the Dutch book in Adversarial Offer with Opt-Out is the following. The Dutch book arises from a disagreement between CDT on Monday and CDT on Tuesday (cf. the discussion under Section IV.3). If the buyer was able to precommit on Monday to a course of action to be followed on Tuesday, the sure loss of money in Adversarial Offer with Opt-Out would be avoided. However, as noted in Section III, we assume that the agent cannot in fact precommit. We find this assumption realistic—certainly humans are unable to arbitrarily precommit to courses of action.

Instead of assuming the ability to precommit, we can modify CDT itself to on Tuesday follow the policy that it would have precommitted to on Monday. Let us refer to this idea as *policy-CDT*, since it asks the agent to evaluate entire policies all at once.[11] In Adversarial Offer with Opt-Out, there are four possible policies: opt out, buy B_1, buy B_2, and buy nothing (where the latter three possibilities include declining the opt-out offer). When considering these policies (*ex ante*), *buy nothing* dominates *opt out*. Hence, policy-CDT will decline the opt-out offer and thereby avoid the Dutch book. Note, however, that such a modification of CDT will make no difference to the choices it prescribes in Adversarial Offer, which has only one decision point. Hence, it will still lose money in expectation.

While this appears to be a promising approach, it is nontrivial to flesh out, because on other examples it is less clear what policy-CDT should prescribe. For illustration, consider the following interpretation of policy-CDT: follow

[10] We could also use the scenario from Section IV.1, in which the predictor can observe whether the buyer randomises and then removes the money from the boxes if she observes randomisation.

[11] Policy-CDT resembles Fisher's (2020) disposition-based decision theory. Compare Meacham (2010) for a discussion of explicit precommitment. Similarly, Gauthier (1989) has argued for evaluating 'plans' not decisions in Newcomb-like problems (without basing this argument on any particular theory like CDT or EDT). Yet another formal treatment is given by Everitt *et al.* (2015). A few authors have also proposed policy versions of other, more EDT-like decision theories (Drescher 2006, section 6.2; Yudkowsky and Soares 2018, section 4). The idea of precommitments has also been discussed outside the literature on Newcomb-like problems. The best-known account is perhaps McClennen's (1990) notion of resolute choice, which has also been discussed by Greene (2018, section 3.2.2) in the context of Newcomb's problem.

the policy to which CDT would like to commit *ex ante*, where *ex ante* refers to some point in time before the first decision of the scenario. Now, let us consider the following scenario based on Newcomb's problem. On Thursday, the agent makes an inconsequential decision, such as whether to eat a peppermint. (We might imagine that at this point, the agent does not yet know what will happen on later days, i.e. that the following only happens with small subjective probability.) On Friday, the predictor observes the agent—again, we could imagine that she uses fMRI. On Saturday, the agent faces Newcomb's problem, where the prediction is based only on the data from Friday's observations. As usual, we imagine that the agent cannot *in fact* precommit on Thursday. The ex-ante-commitment interpretation of policy-CDT would recommend one-boxing on Saturday. Note that this decision hinges on the presence of the decision point on Thursday. To the causal decision theorist, this may be unacceptable, given that adding the peppermint decision is such a minor modification of Newcomb's problem.[12] Relatedly, imagine that in Adversarial Offer with Opt-Out, the opt-out decision on Monday is not made by the buyer himself but by his wife. The buyer and his wife have a shared bank account, have the same beliefs and both use the same decision theory. If they both use standard CDT, then the argument of Section III applies and the wife pays $0.20 in order for her husband to not play Adversarial Offer on Tuesday. It is unclear to us how policy-CDT should deal with this problem.

Imprecise probabilities. Many other ways of modifying CDT are worth considering. For instance, in Adversarial Offer, it may be unrealistic for the buyer to form a single probability distribution over box contents. Instead, he may consider multiple different probability distributions, including one under which box B_1 is probably empty and one under which box B_2 is probably empty. He could then evaluate each option pessimistically, i.e. w.r.t. the probability distribution that is worst under that option. Such a version of CDT would prescribe declining to buy a box. At the same time, it would recommend two-boxing in Newcomb's problem and more generally obey the causal dominance principle. For a discussion of this maxmin criterion for choice under multiple probability distributions, see Gilboa and Schmeidler (1989) and in particular the game-theoretic interpretation of Grünwald and Halpern (2011). A more general discussion of how to use sets of probability distributions (possibly in combination with decision rules other than the maxmin criterion) is offered by Bradley (2012). In our setting, B_1 are B_2 are, roughly, complementary bets in the causalist's beliefs. In all worlds in which B_i is empty, B_{3-i} is full. As discussed by Bradley, it has been argued that a rational agent should

[12] In McClennen's (1990) terminology, policy-CDT violates the *separability condition*: what policy-CDT recommends in a particular situation (like Newcomb's problem) depends on what decision situations the agent has previously been in (such as deciding whether to eat a peppermint), as well as what situations the agent could have been in.

always accept one of a pair of complementary bets. Indeed, expected utility maximisation for a single probability distribution satisfies this complementarity criterion—to the causalist's detriment in Adversarial Offer. Bradley (2012) argues that in general, an agent with imprecise probabilities should not satisfy the complementarity criterion, and that this allows him to avoid Dutch books—though, of course, he considers Dutch books of a very different type.

IV.5 Abandoning the core of CDT

Finally, one may view at least one of the scenarios in this paper as a persuasive argument against the core of CDT and in particular two-boxing. We should then adopt a theory that one-boxes in Newcomb's problem. EDT is the obvious candidate. Of course, even one-boxers have criticised EDT by alleging irrational prescriptions in other cases—such as the Smoking lesion (Ahmed 2014, section 4.1–4.3) or cases of dynamic inconsistency like Newomb's problem with transparent boxes (and the problems listed in footnote 9). In response, various other one-boxing theories have been developed (see, e.g. Gauthier 1989; Spohn 2012; Poellinger 2013; Levinstein and Soares 2020). Like EDT, these theories (as we understand them) recommend not buying a box in Adversarial Offer and thus avoid the loss of money. They could therefore claim our scenarios as supporting their theories relative to CDT.

V CONCLUSION

We have presented Adversarial Offer as a decision problem in which an orthodox causal decision theorist voluntarily accepts the loss of money in expectation. We have also provided a dynamic extension of the scenario, in which the causal decision theorist anticipates his own *expected* loss of money and thus accepts a smaller *certain* loss to avoid facing the scenario in the first place.

We then discussed various responses available to a causal decision theorist. In this discussion, we aimed to state and analyse many possible responses, including ones we do not find plausible ourselves—such as accepting the expected loss of money. Of course, there may always be alternative responses or lines of argument that we have missed. Nevertheless, we conclude from this discussion that Adversarial Offer (and to some extent its dynamic extension) is a devastating counterexample to orthodox, two-boxing CDT. To avoid the problem, we need to substantially modify CDT, e.g. by using imprecise probabilities or even abandoning the causal dominance principle in its usual form.[13]

[13] This work was supported by the National Science Foundation under Award IIS-1814056. We thank Jesse Clifton, Sven Neth, Johannes Treutlein, and our anonymous referees for helpful comments.

REFERENCES

Ahmed, A. (2012) 'Push the Button', *Philosophy of Science*, 79: 386–95.

——— (2014) *Evidence, Decision and Causality*. Cambridge: CUP.

——— (2020) 'Sequential Choice and the Agent's Perspective', https://www.academia.edu/36270656/Sequential_Choice_and_the_Agents_Perspective. Accessed 27 November 2020.

Ahmed, A. and Price, H. (2012) 'Arntzenius on 'Why ain'cha rich?'', *Erkenntnis*, 77: 15–30.

Almond, P. (2010) *On Causation and Correlation Part 1: Evidential Decision Theory is Correct*, https://casparoesterheld.files.wordpress.com/2016/12/almond_edt_1.pdf. Accessed 27 Nov 2020.

Arntzenius, F. (2008) 'No Regrets, or: Edith Piaf Revamps Decision Theory', *Erkenntnis*, 68: 277–97.

Barnes, R. E. (1997) 'Rationality, Dispositions, and the Newcomb Paradox', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 88: 1–28.

Bradley, S. (2012) 'Dutch Book Arguments and Imprecise Probabilities', in D. Dieks *et al.* (eds) *Probabilities, Laws, and Structures*, vol. 3 of The Philosophy of Science in a European Perspective, pp. 3–17. Dordrecht: Springer.

Cavalcanti, E. G. (2010) 'Causation, Decision Theory, and Bell's Theorem: A Quantum Analogue of the Newcomb Problem', *The British Journal for the Philosophy of Science*, 61: 569–97.

Conitzer, V. (2015) 'A Dutch Book Against Sleeping Beauties who are Evidential Decision Theorists', *Synthese*, 192: 2887–99.

Drescher, G. L. (2006) *Good and Real – Demystifying Paradoxes from Physics to Ethics*. Cambridge, MA: MIT Press.

Everitt, T., Leike, J. and Hutter, M. (2015) 'Sequential Extensions of Causal and Evidential Decision Theory', in T. Walsh (ed.) *Algorithmic Decision Theory: 4th International Conference, ADT 2015, Lexington, KY, USA, September 27–30, 2015, Proceedings*, pp. 205–21. Springer.

Farber, N. (2015) 'The Surprising Psychology of Rock-Paper-Scissors', https://www.psychologytoday.com/us/blog/the-blame-game/201504/the-surprising-psychology-rock-paper-scissors. Accessed 27 Nov 2020.

Fisher, J. C. (2020) 'Disposition-Based Decision Theory', https://casparoesterheld.files.wordpress.com/2019/02/dbdt.pdf. Accessed 27 Nov 2020.

Gauthier, D. (1989) 'In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality)', in *Proceedings of the Aristotelian Society, New Series, 1988–1989*, 89: 179–94. London: Oxford University Press.

Gibbard, A. and Harper, W. L. (1981) 'Counterfactuals and Two Kinds of Expected Utility', in W. L. Harper, R. Stalnaker and G. Pearce (eds) *IFS: Conditionals, Belief, Decision, Chance and Time*, vol. 15 of The University of Western Ontario Series in Philosophy of Science. A Series of Books in Philosophy of Science, Methodology, Epistemology, Logic, History of Science, and Related Fields, pp. 153–90. Dordrecht: Springer.

Gilboa, I. and Schmeidler, D. (1989) 'Maxmin Expected Utility with Non-Unique Prior', *Journal of Mathematical Economics*, 18: 141–53.

Greene, P. (2018) 'Success-First Decision Theories', in A. Ahmed (ed.) *Newcomb's Problem*, pp. 115–37. Cambridge: CUP .

Grünwald, P. D. and Halpern, J. Y. (2011) 'Making Decisions Using Sets of Probabilities: Updating, Time Consistency, and Calibration', *Journal of Artificial Intelligence Research*, 42: 393–426.

Hájek, A. (2009) 'Dutch Book Arguments', in *The Handbook of Rational and Social Choice*, chap. 7. Oxford: OUP.

Joyce, J. M. (1999) *The Foundations of Causal Decision Theory*, Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge: CUP.

——— (2012) 'Regret and Instability in Causal Decision Theory', *Synthese*, 187: 123–45.

Levinstein, B. A. and Soares, N. (2020) 'Cheating Death in Damascus', *The Journal of Philosophy*, 117: 237–66.

McClennen, E. F. (1990) *Rationality and Dynamic Choice. Foundational Explorations*, Cambridge: CUP.

Meacham, C. J. G. (2010) 'Binding and its Consequences', *Philosophical Studies*, 149: 49–71.

Nozick, R. (1969) 'Newcomb's Problem and Two Principles of Choice', in N. Rescher *et al.* (ed.) *Essays in Honor of Carl G. Hempel*, pp. 114–46. Dordrecht: Springer.
Oesterheld, C. (2019) 'Approval-Directed Agency and the Decision theory of Newcomb-Like Problems', *Synthese*, doi:10.1007/s11229-019-02148-2.
Parfit, D. (1984) *Reasons and Persons*. Oxford: OUP.
Poellinger, R. (2013) 'Unboxing the Concepts in Newcomb's Paradox: Causation, Prediction, Decision'. http://philsci-archive.pitt.edu/9887/7/newcomb_in_ckps.pdf. Accessed 27 Nov 2020.
Rabinowicz, W. (2000) 'Money Pump with Foresight', in M. J. Almeida (ed.) *Imperceptible Harms and Benefits*, pp. 123–54. Dordrecht: Springer.
——— (2008) 'Pragmatic Arguments for Rationality Constraints', in M.-C. Galavotti, R. Scazzieri and P. Suppes (eds) *Reasoning, Rationality and Probability*, pp. 139–63. CSLI Publications. https://lup.lub.lu.se/search/publication/737996. Accessed 6 January 2021.
Skalse, J. (2021) 'A General Counterexample to any Decision Theory and Some Responses', https://arxiv.org/abs/2101.00280v1.
Skyrms, B. (1993) 'A Mistake in Dynamic Coherence Arguments?', *Philosophy of Science*, 60: 320–28.
Soares, N. and Fallenstein, B. (2014) 'Toward Idealized Decision Theory', Tech. Rep. 2014-7, Machine Intelligence Research Institute.
Spencer, J. and Wells, I. (2017) 'Why Take Both Boxes?', *Philosophy and Phenomenological Research*, 99: 27–48.
Spohn, W. (2012) 'Reversing 30 Years of Discussion: why Causal Decision Theorists Should One-Box', *Synthese*, 187: 95–122.
Steele, K. and Stefánsson, H. O. (2016) 'Decision Theory', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edn.
S. Vineberg (2016) 'Dutch Book Arguments', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2016 edn. Stanford, CA: Metaphysics Research Lab, Stanford University.
Weirich, P. (1985) 'Decision Instability', *Australasian Journal of Philosophy*, 63: 465–72.
——— (2016) 'Causal Decision Theory', in *The Stanford Encyclopedia of Philosophy*. Spring 2016 edn. Stanford, CA: Stanford University.
Yudkowsky, E. (2010) 'Timeless Decision Theory', http://intelligence.org/files/TDT.pdf. Accessed 27 Nov 2020.
Yudkowsky, E. and Soares, N. (2018) 'Functional Decision Theory: A New Theory of Instrumental Rationality', https://arxiv.org/abs/1710.05060v2. Accessed 6 January 2021.

Duke University, NC, USA