Extracting Money from Causal Decision Theorists

Caspar Oesterheld[*] , **Vincent Conitzer**
Duke University, Department of Computer Science
{ocaspar, conitzer}@cs.duke.edu

Abstract

Newcomb's problem has spawned a debate about which variant of expected utility maximization (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.

1 Introduction

In Newcomb's problem [Nozick, 1969; Ahmed, 2014], 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?

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 one box. Hence, I should choose both boxes.

Evidential decision theory (EDT), on the other hand, recommends that the agent reason as follows: if I choose one box, then in all likelihood the being predicted that I would choose one box, 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, then I can expect to walk away with (close to) $1,000. Hence, I should choose one box.

While Newcomb's problem itself is far-fetched, it has been argued that the difference between CDT and EDT matters in various game-theoretic settings. First, it has been pointed out that Newcomb's problem closely resembles playing a Prisoner's Dilemma against a similar opponent [Brams, 1975; Lewis, 1979]. Whereas CDT recommends defecting even in a Prisoner's Dilemma against an exact copy, EDT recommends cooperating when facing sufficiently similar opponents. What degree of similarity is required and whether EDT and CDT ever come apart in this way *in practice* has been the subject of much discussion [Ahmed, 2014, Section 4.6]. A common view is that CDT and EDT come apart only under fairly specific circumstances that do not include most human interactions. Still, Hofstadter [1983], for example, argues that "superrational" humans should cooperate with each other in a real-world one-shot Prisoner's Dilemma, reasoning in a way that resembles the reasoning done under EDT. Economists have usually been somewhat dismissive of such ideas, sometimes referring to them as "(quasi-)magical thinking" when trying to explain observed human behavior [Shafir and Tversky, 1992; Masel, 2007; Daley and Sadowski, 2017]. Indeed, standard game-theoretic solution concepts are closely related to ratificationism [Joyce and Gibbard, 1998, Section 5], a variant of CDT which we will revisit later in this paper (see Part 3 of Section 4).

Second, even if the players of a game are dissimilar, it is in the nature of strategic interactions that each player's behavior is predicted by the other players. Newcomb's problem itself (as well as the ADVERSARIAL OFFER discussed in this paper) differs from strategic interactions as they are usually considered in game theory in its asymmetry. However, one might still expect that CDT and EDT offer different perspectives on how rational agents should deal with the mutual prediction inherent in strategic interactions [Gauthier, 1989, Section XI].

Third, CDT and EDT differ in their treatment of situations with imperfect recall. Seminal discussions of such games are due to Piccione and Rubinstein [1997] and Aumann *et al.* [1997] [cf. Bostrom, 2010, for an overview from a philosopher's perspective]. While these problems originally were not associated with Newcomb's problem, the relevance of different decision theories in this context has been pointed out by Briggs [2010] [cf. Armstrong, 2011; Schwarz, 2015; Conitzer, 2015].

The importance of the differences in decision theory is amplified if, instead of humans, we consider artificial agents. After all, it is common that multiple copies of a software sys-

[*]Contact Author

tem are deployed and other parties are often able to obtain the source code of a system to analyze or predict its behavior. As some software systems choose more autonomously, we might expect their behavior will be (approximately) describable by CDT or EDT (or yet some other theory). If either of these theories has serious flaws, we might worry that if a system implements the wrong theory, it will make unexpected, suboptimal choices in some scenarios. Such scenarios might arise naturally, e.g., as many copies of a system are deployed. We might also worry about adversarial problems like the one in this paper.

One argument against CDT is that 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).[1] Constructing such a scenario from Newcomb's problem is non-trivial. For example, in Newcomb's problem, a causal decision theorist may realize 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 2). 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 3). Finally, we discuss the implications of the existence of such scenarios (Section 4).

[1]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, 2018, for a harder-to-defuse version of this point]. 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, e.g., 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.

2 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. No randomization device is available to the buyer (or at least no randomization device that is not predictable to the seller).[2]

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

$$\begin{aligned} & \$1 - P(\$3 \text{ in } B_i \mid \text{buyer chooses } B_i) \cdot \$3 \\ = \; & \$1 - 0.25 \cdot \$3 \\ = \; & \$0.25. \end{aligned}$$

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; Soares and Levinstein, 2017].

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

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

where $P(\$3 \text{ in box } B_i)$ 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 $P(\$3 \text{ in box } B_i) = p_0 + p_{3-i}$ for $i = 1, 2$, it is $P(\$3 \text{ in box } B_i) \geq \frac{1}{2}$ for at least one $i \in \{1, 2\}$. Thus, the expected utility in eq. 1 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 realize that the box he bought is empty

[2]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.

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].

3 A diachronic Dutch book against causal decision theory

ADVERSARIAL OFFER results in a loss *in expectation* for the causal decision theorist. It is natural to ask whether it is possible to set up the scenario so that the causal decision theorist ends up with a *sure* loss; effectively, a Dutch book. 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 perfectly predict the buyer 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].[3] To get our Dutch book to work, we add another choice *before* ADVERSARIAL OFFER.

> 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.

4 Discussion

We differentiate four types of responses to these scenarios available to supporters of causal decision theory:

1. They could claim that these scenarios are irrelevant for evaluating decision theories, in the sense that they are impossible to set up or otherwise out of scope, and therefore unpersuasive.
2. They could concede that these scenarios are relevant for evaluating decision theories, but claim that CDT's recommendations in them are acceptable.
3. 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 the causal dominance principle.
4. They could concede that these scenarios show that the very core of CDT (two-boxing and the causal dominance principle) is implausible.

We will discuss these options in turn.

1 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 defense of CDT. In particular, a causal decision theorist might claim that sufficiently accurate prediction of a CDT agent is simply impossible.[4] 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

[3] 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.

[4] For a general discussion of such unpredictability claims in defense of CDT, see Ahmed [2014, Chapter 8].

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 the ADVERSARIAL OFFER – most people follow exploitable patterns in what moves they select [see, e.g., Farber, 2015, and references therein].[5] 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 defense for the supporter of causal decision theory 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 CDT also requires being unpredictable to the seller, either by succeeding at out-thinking the seller sufficiently often, or by acting sufficiently randomly.

Is it reasonable to consider this a requirement of acting according to CDT? CDT does not suggest any strict preference for choosing randomly across options, as opposed to just deterministically choosing one of the options that is best according to the buyer's beliefs. Hence, the unpredictability would have to emerge from the buyer attempting to out-think the seller. But it does not seem that this is always an attainable goal. 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 behavior 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 realized 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.].

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 the ADVERSARIAL OFFER) resemble Newcomb's problem – the problem that has led to the development of CDT in the first place. 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 behavior is determined by a computer program) facing a wide range of scenarios including the ones given in this paper.

2 If our scenarios are within the scope of causal decision theory, then the supporter of causal decision theory has to contend with the fact that one can extract expected money from, and even Dutch-book, CDT agents in them. 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, e.g., Vineberg [2016] or Hájek [2009]. Note, though, that some of the most influential arguments in favor of expected utility maximization (EUM) – of which CDT is a refinement – are Dutch books. Of course, one may use different arguments to justify EUM. But it would seem odd to follow Dutch-book arguments to EUM 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.[6] 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 observes the agent (perhaps using 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 [Gibbard and Harper, 1981, Section 10; also discussed by Gauthier, 1989; Drescher, 2006, Section 6.2; Arntzenius, 2008, Section 7; Meacham, 2010, Section 3.2.2]. 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.[7] 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.

[5] There are multiple rock-paper-scissors bots available online which attempt to predict their opponent's future moves based on past moves (using data from other players). As of July 2019, the bot at http://www.essentially.net/rsp/ has reportedly played about 2 million rounds and won 57% more often than it lost.

[6] For a discussion of similar arguments about other diachronic Dutch books, see, e.g., Rabinowicz [2008].

[7] Parfit's (1984) hitchhiker [Barnes, 1997], XOR Blackmail [Soares and Levinstein, 2017, Section 2] and Yankees vs. Red Sox [Arntzenius, 2008, pp. 22-23; Ahmed and Price, 2012] 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.

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 the 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. Conversely, defenses of dynamic inconsistencies [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 defenses 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. 22-23, Section 4]. In sharp contrast, the ADVERSARIAL OFFER WITH OPT-OUT lacks any such event between the decision points. The difference in 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, we agree that 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 [Steele and Stefánsson, 2016, Section 6]. Nevertheless, we do not see any clear path by which such research would justify CDT's recommendations in the 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.

3 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]. 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 the 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 you will take box B_i, it would be better to buy the other box B_{3-i}.

The ratificationist may respond by claiming that unpredictable randomization 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 true random number generator (TRNG) (e.g., one based on nuclear decay) is in fact available, this does not settle the issue. For example, consider a variant of the 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 output of the TRNG. In this ANTI-RANDOMIZATION ADVERSARIAL OFFER, again no option is ratifiable: under the beliefs that would result from knowing that you will make different choices depending on the TRNG's output (and therefore choose a box with some positive probability), you would rather not pick any box. To circumvent this example, the ratificationist could argue that the decision maker should be able to randomize in such a way that *whether* he is randomizing 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 **1**.

A different strategy for modifying CDT to avoid the Dutch book in the 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 **2**). A tempting possibility is to modify CDT so that it considers all decisions to be made at once. That is, such a version of CDT – let us refer to it as *policy-CDT* – prescribes that one decide on one's general *policy* all at once.[8] In the ADVERSARIAL OFFER WITH OPT-OUT, there are four possible policies: opt out, buy B_1, buy B_2, and buy nothing (where the last three possibilities include declining the opt-out offer). When considering these policies, *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 the policy to which CDT would like to *commit* ex ante, where "ex

[8] Policy-CDT resembles Fisher's [nd] 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). 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].

ante" refers to some point in time before the first decision of the scenario. Now, let us consider a version of Newcomb's problem which is supplemented by another trivial and unrelated decision – say, whether to eat a peppermint – that takes place when the agent still has a causal influence over the prediction. Then the ex-ante-commitment interpretation of policy-CDT would recommend one-boxing. To the causal decision theorist, this may be unacceptable, especially given that adding the peppermint decision is such a minor modification of Newcomb's problem. Perhaps there is a way to define policy-CDT that avoids such dependence on irrelevant decisions while also prescribing two-boxing, but it is not immediately obvious how to do so.

Many other ways of modifying CDT are worth considering. For instance, in the 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, e.g., Gilboa and Schmeidler [1989] and in particular game-theoretic interpretations such as that of Grünwald and Halpern [2011]. A more general discussion of how using sets of probability distributions (while potentially 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 accept one of a pair of complementary bets. Indeed, expected utility maximization for a single probability distribution satisfies this complementarity criterion – to the causalist's detriment in the 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.

4 Finally, one may view at least one of the scenarios in this paper as supporting a persuasive argument against the very core of CDT. EDT is the obvious alternative. However, depending on how problematic we find EDT's 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 7) – we may also look to various other decision theories that have been proposed [Gauthier, 1989; Spohn, 2012; Poellinger, 2013; Soares and Levinstein, 2017].

Acknowledgements

We thank Johannes Treutlein and Jesse Clifton for comments and discussions.

References

[Ahmed and Price, 2012] Arif Ahmed and Huw Price. Arntzenius on 'why ain'cha rich?'. *Erkenntnis*, 77(1):15–30, 7 2012.

[Ahmed, 2012] Arif Ahmed. Push the button. *Philosophy of Science*, 79(3):386–395, 7 2012.

[Ahmed, 2014] Arif Ahmed. *Evidence, Decision and Causality*. Cambridge University Press, 2014.

[Almond, 2010] Paul Almond. On causation and correlation part 1: Evidential decision theory is correct, 9 2010.

[Armstrong, 2011] Stuart Armstrong. Anthropic decision theory, Nov 2011.

[Arntzenius, 2008] Frank Arntzenius. No regrets, or: Edith piaf revamps decision theory. *Erkenntnis*, 68(2):277–297, 2008.

[Aumann *et al.*, 1997] Robert J. Aumann, Sergiu Hart, and Motty Perry. The absent-minded driver. *Games and Economic Behavior*, 20:102–116, 1997.

[Barnes, 1997] R. Eric Barnes. Rationality, dispositions, and the newcomb paradox. *Philosophical Studies: An International Journal for Philosophy in the AnalyticTradition*, 88(1):1–28, 10 1997.

[Bostrom, 2010] Nick Bostrom. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Studies in Philosophy. Routledge, 2010.

[Bradley, 2012] Seamus Bradley. Dutch book arguments and imprecise probabilities. In D. Dieks, W. Gonzalez, S. Hartmann, M. Stöltzner, and M. Weber, editors, *Probabilities, Laws, and Structures*, volume 3 of *The Philosophy of Science in a European Perspective*, pages 3–17. Springer, Dordrecht, 2012.

[Brams, 1975] Steven J. Brams. Newcomb's problem and prisoners' dilemma. *The Journal of Conflict Resolution*, 19(4):596–612, 12 1975.

[Briggs, 2010] Rachael Briggs. Putting a value on beauty. volume 3 of *Oxford Studies in Epistemology*, pages 3–34. Oxford University Press, 2010.

[Cavalcanti, 2010] Eric G. Cavalcanti. Causation, decision theory, and bell's theorem: A quantum analogue of the newcomb problem. *The British Journal for the Philosophy of Science*, 61(3):569–597, 9 2010.

[Conitzer, 2015] Vincent Conitzer. A dutch book against sleeping beauties who are evidential decision theorists. *Synthese*, 192(9):2887–2899, 10 2015.

[Daley and Sadowski, 2017] Brendan Daley and Philipp Sadowski. Magical thinking: A representation result. *Theoretical Economics*, 12:909–956, 2017.

[Drescher, 2006] Gary L. Drescher. *Good and Real – Demystifying Paradoxes from Physics to Ethics*. MIT Press, 2006.

[Farber, 2015] Neil Farber. The surprising psychology of rock-paper-scissors, 4 2015.

[Fisher, nd] Justin C. Fisher. Disposition-based decision theory. n.d.

[Gauthier, 1989] David Gauthier. In the neighbourhood of the newcomb-predictor (reflections on rationality). In *Proceedings of the Aristotelian Society, New Series, 1988–1989*, volume 89, pages 179–194. 1989.

[Gibbard and Harper, 1981] Allan Gibbard and William L. Harper. Counterfactuals and two kinds of expected utility. In William L. Harper, Robert Stalnaker, and Glenn Pearce, editors, *Ifs. Conditionals, Belief, Decision, Chance and Time*, volume 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*, pages 153–190. Springer, 1981.

[Gilboa and Schmeidler, 1989] Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18:141–153, 1989.

[Grünwald and Halpern, 2011] Peter D. Grünwald and Joseph Y. Halpern. Making decisions using sets of probabilities: Updating, time consistency, and calibration. *Journal of Artificial Intelligence Research*, 42, 2011.

[Hofstadter, 1983] Douglas Hofstadter. Dilemmas for superrational thinkers, leading up to a luring lottery. *Scientific American*, 248(6), Jun 1983.

[Hájek, 2009] Alan Hájek. Dutch book arguments. In *The Handbook of Rational and Social Choice*, chapter 7. Oxford University Press, 2009.

[Joyce and Gibbard, 1998] James M. Joyce and Allan Gibbard. Causal decision theory. In *Handbook of Utility Theory, Volume 1: Principles.*, chapter 13, pages 627–666. Kluwer, 1998.

[Joyce, 1999] James M. Joyce. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge University Press, 1999.

[Joyce, 2012] James M. Joyce. Regret and instability in causal decision theory. *Synthese*, 187:123–145, 2012.

[Lewis, 1979] David Lewis. Prisoners' dilemma is a newcomb problem. *Philosophy & Public Affairs*, 8(3):235–240, 1979.

[Masel, 2007] Joanna Masel. A bayesian model of quasi-magical thinking can explain observed cooperation in the public good game. *Journal of Economic Behavior & Organization*, 64(2):216–231, 10 2007.

[Meacham, 2010] Christopher J. G. Meacham. Binding and its consequences. *Philosophical Studies*, 149(1):49–71, 5 2010.

[Nozick, 1969] Robert Nozick. Newcomb's problem and two principles of choice. In Nicholas Rescher et al., editor, *Essays in Honor of Carl G. Hempel*, pages 114–146. Springer, 1969.

[Oesterheld, 2019] Caspar Oesterheld. Approval-directed agency and the decision theory of newcomb-like problems. *Synthese*, 2019.

[Parfit, 1984] Derek Parfit. *Reasons and Persons*. Oxford University Press, 1984.

[Piccione and Rubinstein, 1997] Michele Piccione and Ariel Rubinstein. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior*, 20:3–24, 1997.

[Poellinger, 2013] Roland Poellinger. Unboxing the concepts in newcomb's paradox: Causation, prediction, decision. 2013.

[Rabinowicz, 2000] Wlodek Rabinowicz. Money pump with foresight. In Michael J. Almeida, editor, *Imperceptible Harms and Benefits*, pages 123–154. Springer, 2000.

[Rabinowicz, 2008] Wlodek Rabinowicz. *Pragmatic Arguments for Rationality Constraints*, pages 139–163. Reasoning, Rationality and Probability. CSLI Publications, 2008.

[Schwarz, 2015] Wolfgang Schwarz. Lost memories and useless coins: revisiting the absentminded driver. *Synthese*, 192:3011–3036, 2015.

[Shafir and Tversky, 1992] Eldar Shafir and Amos Tversky. Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24(4):449–474, 1992.

[Skalse, 2018] Joar Skalse. A counterexample to perfect decision theories and a possible response. 2018.

[Skyrms, 1993] Brian Skyrms. A mistake in dynamic coherence arguments? *Philosophy of Science*, 60:320–328, 6 1993.

[Soares and Fallenstein, 2014] Nate Soares and Benja Fallenstein. Toward idealized decision theory. Technical Report 2014-7, Machine Intelligence Research Institute, 2014.

[Soares and Levinstein, 2017] Nate Soares and Benjamin A. Levinstein. Cheating death in damascus. In *Formal Epistemology Workshop (FEW) 2017*, University of Washington, Seattle, USA, 5 2017.

[Spencer and Wells, 2017] Jack Spencer and Ian Wells. Why take both boxes? *Philosophy and Phenomenological Research*, 2017.

[Spohn, 2012] Wolfgang Spohn. Reversing 30 years of discussion: why causal decision theorists should one-box. *Synthese*, 187(1):95–122, 2012.

[Steele and Stefánsson, 2016] Katie Steele and H. Orri Stefánsson. Decision theory. 2016.

[Vineberg, 2016] Susan Vineberg. Dutch book arguments. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2016 edition, 2016.

[Weirich, 1985] Paul Weirich. Decision instability. *Australasian Journal of Philosophy*, 63(4):465–472, 1985.

[Weirich, 2016] Paul Weirich. Causal decision theory. In *The Stanford Encyclopedia of Philosophy*. Spring 2016 edition, 2016.

[Yudkowsky and Soares, 2018] Eliezer Yudkowsky and Nate Soares. Functional decision theory: A new theory of instrumental rationality, 5 2018.

[Yudkowsky, 2010] Eliezer Yudkowsky. Timeless decision theory, 2010.