Extracting Money from Causal Decision Theorists

Caspar Oesterheld and Vincent Conitzer

Duke University

January 10, 2019

Abstract

In Newcomb's problem, causal decision theorists walk away with less money than evidential decision theorists. However, this argument has proven unpersuasive in the debate about whether causal or evidential decision theory (CDT or EDT) should be preferred. In this paper, we provide two scenarios in which CDT voluntarily loses money to another agent. Arguably, such scenarios are more decisive. In our first such scenario, the agent faces a single choice and following CDT's recommendation yields a loss of money in expectation. The second extends the first to yield a diachronic Dutch book of CDT.

Keywords: Newcomb's problem, causal decision theory, evidential decision theory, Dutch Book Arguments

1 Introduction

In Newcomb's problem (Nozick, 1969; see Ahmed, 2014, for a general overview of the literature on Newcomb's problem and the foundations of decision theory), a "being" puts non-negative amounts of money in two boxes, A and B. An agent is then asked to choose between receiving the contents of both boxes, or of box B only. However, before putting money into the boxes, the being predicts whether the agent will choose one or two boxes. In the former case, the being will put \$1,000,000 in box B; in the latter case, the being will put nothing in box B. The being puts \$1,000 in the right box in any case. The being's predictions are uncannily accurate. What should the agent do?

Causal decision theory (CDT) endorses the following line of reasoning: 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, endorses the following line of reasoning: 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.

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. It would be more convincing to provide 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 (sect. 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 (sect. 3). Finally, we discuss the implications of the existence of such scenarios (sect. 4).

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]

[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 argument). However, such a setup does not allow one to generate a generic scenario in which an agent loses money in spite of having the option to walk away losing nothing.

Furthermore, scenarios with expected loss appear significantly more problematic for pragmatic reasons: if – as in the scenarios that we will provide here – the causal decision theorist's loss is someone else's profit, this generates a significant incentive to place him in such a situation.

[2] This decision problem resembles the widely discussed Death in Damascus scenario (introduced to the decision theory literature by Gibbard and Harper, 1981, sect. 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.

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; Soares and Levinstein, 2017).

In contrast, CDT prescribes that the buyer buy one of the two boxes, for the following reasons. 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 $P(\text{money in box } B_i) \geq \frac{1}{2}$ 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 realize 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).

It is informative to consider this in light of van Fraassen's (1984) belief reflection principle, which states: If tomorrow I will be better informed than today, and my belief in event E will be p regardless of what information I receive between now and then, then my belief in E should already be p now. In our example, consider the event $E_{-\$1}$: "At the end I will have lost \$1 overall." If the buyer takes a box, then *regardless of which box it is*, afterwards his subjective probability on $E_{-\$1}$ will be 0.75. Now, consider the buyer's situation before choosing a box. He can conclude that he will buy a box, by the same reasoning we did earlier. Which box that is may be information that will come to him only later, but because his belief in $E_{-\$1}$ will be the same regardless of which box it is, there is a strong argument to be made that the reflection principle should apply, and hence he should currently believe that there is a 75% chance he will

end up having lost \$1. By entirely similar reasoning, he should currently believe there is a 25% chance that he will end up having gained \$2. Hence, even before making a decision, his expected value from the game as a whole is −\$.25. While a causal decision theorist's beliefs may satisfy this implication of the reflection principle on some level, the beliefs that inform his decisions do not; otherwise, he would not buy a box in the first place.

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. 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: afterwards, allow the buyer to reverse his decision for a small fee (ending up without any box and having lost only the fee). However, this does not work: if we add this option, then he will anticipate eventually undoing his choice and therefore not buy a box in the first place (Ahmed, 2014, sect. 3.2; though cf. Rabinowicz, 2000).[3] Instead, 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, the buyer cannot, on Monday, commit himself to a course of action to follow on Tuesday.

First, 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.

Will the causal decision theorist accept the Monday offer? Here is an argument that he will.

[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.

1. The expected utility that he assigns to (not) taking the opt-out offer when deciding whether to take it should be equal to his perceived expected utility on Monday night, after (not) taking the opt-out offer.

2. His perceived expected utility on Monday night is higher after taking the opt-out offer than it is after declining it.

$\Longrightarrow$ Therefore, he should take the opt-out offer.

Let us first argue for 1. In some decision scenarios, of course, the expected utility that the causal decision theorist assigns to a choice while contemplating that choice is different from his perceived expected utility immediately after that choice. For example, this is the case in Adversarial Offer; as we argued, when making the choice, he evaluates the expected utility for at least one box as positive, while immediately after choosing that box, he perceives his expected utility to be negative. But this is only because his choice provides evidence about the state of the world (the content of the boxes) that is unrelated to the causal effects of the choice. In contrast, it does not seem that his decision about whether to take the opt-out offer provides him with any such evidence about the state of the world – in particular about the content of the boxes, what he would do on Tuesday, etc. Hence, the beliefs that he used while evaluating the decision to (not) opt out should be the same as the beliefs after (not) opting out.

Now, let us argue for 2. Since the buyer doesn't face a decision on Monday night, his perceived utility is informed by his "regular" beliefs (as opposed beliefs involving counterfactuals). Hence, conditional on not opting out, the expected value of his final payout as calculated on Monday night should be -$.25. E.g., as we have argued above, this follows immediately if the buyer accepts van Fraassen's belief reflection principle. If he does opt out, then his final payout is certainly -$.2, which is higher. Hence, 2 holds.

Note that if it is known that the buyer uses CDT on Tuesday, then accepting the Monday offer is uncontroversial. For instance, if the agent followed EDT on Monday and CDT on Tuesday, then he would still accept the Monday offer. Similarly, if the seller suspects that the buyer will pick one of the boxes on Tuesday, then she will hope that he rejects the Monday offer. What creates the opportunity for a Dutch book is the prospect of buying a box on Tuesday, 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 in ways that maintain 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 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 $\$1/(\frac{1}{2} - \epsilon)$ 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, 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 ones that is best according to the agent's beliefs. Hence, the unpredictability would have to emerge from the agent attempting to out-think the seller. But it does not

[4] For a discussion of such unpredictability claims in defense of CDT, see Ahmed (2014, ch. 8).

[5] There are multiple rock-paper-scissors bots available online which attempt to predict (using data from other players) future moves based on past moves. As of December 2018, the bot at `http://www.essentially.net/rsp/` has reportedly played about 1.6 million rounds and won 60% more often than it lost.

seem that this is always an attainable goal. For example, imagine that the agent is a deterministic computer program whose source code is known to the seller. Then regardless of how exactly the agent works, its behavior is perfectly predictable by a (computationally sufficiently powerful) seller (cf. Soares and Fallenstein, 2014, sect. 2; Cavalcanti, 2010, sect. 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 artificially 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 Dutch books are conclusive decision-theoretic arguments, see, e.g., Vineberg (2016) or Hájek (2009) (cf. Arkes, Gigerenzer, and Hertwig, 2016, for a psychological perspective). Note, though, that some of the most influential arguments in favor of expected utility maximization (EUM) generally – 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.

We next discuss a different response that is more specific to CDT and scenarios with multiple decisions over time (as in the case of 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.

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

Consider Newcomb's problem. Let t_0 be the time at which the predictor observes the agent in order to make a prediction. Then, before t_0, CDT recommends committing – and if needed paying money to commit – to one-boxing (cf. 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.

Indeed, EDT faces similar apparent inconsistencies over time. For instance, consider a version of Newcomb's problem in which both boxes are transparent (Gibbard and Harper, 1981, sect. 10; also discussed by Meacham, 2010, sect. 3.2.2; Drescher, 2006, sect. 6.2; Arntzenius, 2008, sect. 7). 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.

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, Ahmed (2014, sect. 3.2) gives a defense of one type of dynamic inconsistencies in Newcomb-like problems that focuses on arguing that they do not allow for Dutch Books.

Another reason that CDT's prescriptions in the Adversarial Offer with Opt-Out are more problematic is the following. For CDT in Newcomb's problem and EDT in Newcomb's problem with transparent boxes, there are particular events that split the decision perspectives. For CDT in Newcomb's problem, that event is the loss of causal control at t_0 over the content of box B. For EDT in the Newcomb's problem with transparent boxes, that event is the loss of evidential control (cf. Almond, 2010, sect. 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, sect. 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 housemade – a result of CDT viewing its current choice differently from 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 (cf. Steele and Stefánsson, 2016, sect. 6). Nevertheless, we do not see any clear path by which such research would justify CDT's recommendations

[7] Soares and Fallenstein (2014, sect. 2.1, "Evidential blackmail") and Arntzenius (2008, "Yankees vs. Red Sox") (see Ahmed and Price, 2012, pp. 22-23, sect. 4, for further discussion) describe scenarios that are very similar to Newcomb's problem with transparent boxes.

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, sect. 5.1) a.k.a. sure thing (Gibbard and Harper, 1981, sect. 7) principle (though see Ahmed, 2012, for a general 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, sect. 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 example, 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. This, of course, is a point we have already addressed under **1**: 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 necessarily 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 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. In the Adversarial Offer with Opt-Out, there are four possible policies: {opt out, buy B_1, buy B_2, 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 at first glance 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 ante" refers to some point in time before the first decision of the scenario.[8] 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 theorists, 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. Then 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).

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

[8] Defined in this way, policy-CDT resembles Fisher's (n.d.) disposition-based decision theory. Compare Meacham (2010) for a discussion of explicit precommitment. A few authors have also proposed policy versions of other more EDT-like decision theories (Drescher, 2006, sect. 6.2; Yudkowsky and Soares, 2018, sect. 4).

cases – such as the Smoking lesion (Ahmed, 2014, sect. 4.1–4.3), Conitzer's (2015) modified Sleeping Beauty case, or Evidential Blackmail (Soares and Fallenstein, 2014, sect. 2.1) – we may also look to various other decision theories that have been proposed (see, e.g. Spohn, 2012; Poellinger, 2013; Soares and Levinstein, 2017).

Acknowledgements

We thank Johannes Treutlein for comments and discussions.

References

Ahmed, Arif (2012). "Push the Button". In: *Philosophy of Science* 79.3, pp. 386–395. URL: https://www.jstor.org/stable/10.1086/666065.

— (2014). *Evidence, Decision and Causality*. Cambridge University Press.

Ahmed, Arif and Huw Price (2012). "Arntzenius on 'Why ain'cha rich?'" In: *Erkenntnis* 77.1, pp. 15–30. DOI: 10.1007/s10670-011-9355-2.

Almond, Paul (2010). *On Causation and Correlation Part 1: Evidential decision theory is correct*. URL: https://casparoesterheld.files.wordpress.com/2016/12/almond_edt_1.pdf.

Arkes, Hal R., Gerd Gigerenzer, and Ralph Hertwig (2016). "How Bad Is Incoherence?" In: *Decision* 3.1, pp. 20–39. URL: http://www.spp1516.de/en/Publications/pdfs/Arkes%20et%20al.%202016_How%20Bad%20Is%20Incoherence.pdf.

Arntzenius, Frank (2008). "No Regrets, or: Edith Piaf Revamps Decision Theory". In: *Erkenntnis* 68.2, pp. 277–297. DOI: 10.1007/s10670-007-9084-8.

Cavalcanti, Eric G. (2010). "Causation, Decision Theory, and Bell's Theorem: A Quantum Analogue of the Newcomb Problem". In: *The British Journal for the Philosophy of Science* 61.3, pp. 569–597. DOI: 10.1093/bjps/axp050.

Conitzer, Vincent (2015). "A Dutch book against sleeping beauties who are evidential decision theorists". In: *Synthese* 192.9, pp. 2887–2899.

Drescher, Gary L. (2006). *Good and Real – Demystifying Paradoxes from Physics to Ethics*. MIT Press. URL: https://www.gwern.net/docs/statistics/decision/2006-drescher-goodandreal.pdf.

Farber, Neil (2015). *The Surprising Psychology of Rock-Paper-Scissors*. URL: https://www.psychologytoday.com/us/blog/the-blame-game/201504/the-surprising-psychology-rock-paper-scissors.

Fisher, Justin C. (n.d.). "Disposition-based decision theory". URL: http://www.justin-fisher.com/papers/DBDT.pdf.

Gibbard, Allan and William L. Harper (1981). "Counterfactuals and Two Kinds of Expected Utility". In: *Ifs. Conditionals, Belief, Decision, Chance and Time*. Ed. by William L. Harper, Robert Stalnaker, and Glenn Pearce. Vol. 15. 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. Springer, pp. 153–190. DOI: 10.1007/978-94-009-9117-0_8.

Gilboa, Itzhak and David Schmeidler (1989). "Maxmin expected utility with non-unique prior". In: *Journal of Mathematical Economics* 18, pp. 141–153.

Grünwald, Peter D. and Joseph Y. Halpern (2011). "Making Decisions Using Sets of Probabilities: Updating, Time Consistency, and Calibration". In: *Journal of Artificial Intelligence Research* 42. DOI: 10.1613/jair.3374.

Hájek, Alan (2009). "Dutch Book Arguments". In: *The Handbook of Rational and Social Choice*. Oxford University Press. Chap. 7.

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

— (2012). "Regret and instability in causal decision theory". In: *Synthese* 187, pp. 123–145. DOI: 10.1007/s11229-011-0022-6.

Meacham, Christopher J. G. (2010). "Binding and its consequences". In: *Philosophical Studies* 149.1, pp. 49–71. DOI: 10.1007/s11098-010-9539-7.

Nozick, Robert (1969). "Newcomb's Problem and Two Principles of Choice". In: *Essays in Honor of Carl G. Hempel*. Ed. by Nicholas Rescher et al. Springer, pp. 114–146. URL: http://faculty.arts.ubc.ca/rjohns/nozick_newcomb.pdf.

Poellinger, Roland (2013). "Unboxing the Concepts in Newcomb's Paradox: Causation, Prediction, Decision". URL: http://philsci-archive.pitt.edu/9887/7/newcomb_in_ckps.pdf.

Rabinowicz, Wlodek (2000). "Money Pump with Foresight". In: *Imperceptible Harms and Benefits*. Ed. by Michael J. Almeida. Springer, pp. 123–154.

— (2008). "Pragmatic Arguments for Rationality Constraints". In: ed. by M-C Galavotti, R Scazzieri, and P Suppes. Reasoning, Rationality and Probability. CSLI Publications, pp. 139–163. URL: https://lup.lub.lu.se/search/publication/737996.

Skalse, Joar (2018). "A counterexample to perfect decision theories and a possible response".

Soares, Nate and Benja Fallenstein (2014). *Toward Idealized Decision Theory*. Tech. rep. 2014-7. Machine Intelligence Research Institute. URL: https://arxiv.org/abs/1507.01986.

Soares, Nate and Benjamin A. Levinstein (2017). "Cheating Death in Damascus". In: *Formal Epistemology Workshop (FEW) 2017*. URL: https://intelligence.org/files/DeathInDamascus.pdf.

Spencer, Jack and Ian Wells (2017). "Why Take Both Boxes?" In: *Philosophy and Phenomenological Research*. URL: https://doi.org/10.1111/phpr.12466.

Spohn, Wolfgang (2012). "Reversing 30 years of discussion: why causal decision theorists should one-box". In: *Synthese* 187.1, pp. 95–122.

Steele, Katie and H. Orri Stefánsson (2016). "Decision Theory". In: ed. by Edward N. Zalta.

van Fraassen, Bas C. (1984). "Belief and the Will". In: *The Journal of Philosophy* 81.5, pp. 235–256.

Vineberg, Susan (2016). "Dutch Book Arguments". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2016. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2016/entries/dutch-book/.

Weirich, Paul (1985). "Decision instability". In: *Australasian Journal of Philosophy* 63.4, pp. 465–472.

— (2016). "Causal Decision Theory". In: *The Stanford Encyclopedia of Philosophy*. Spring 2016. URL: https://plato.stanford.edu/archives/spr2016/entries/decision-causal/.

Yudkowsky, Eliezer (2010). *Timeless Decision Theory*. The Singularity Institute. URL: http://intelligence.org/files/TDT.pdf.

Yudkowsky, Eliezer and Nate Soares (2018). *Functional Decision Theory: A New Theory of Instrumental Rationality*. URL: https://arxiv.org/abs/1710.05060v2.