ON DICING WITH DEATH: DEFENDING CAUSAL DECISION THEORY AGAINST UNCANNY CORRELATION

Yucheng Lu[1]

yuchenglu@nyu.edu

ABSTRACT

This dissertation defends Causal Decision Theory(CDT) against a recent (alleged) counterexample. In Dicing with Death (2014), Arif Ahmed devises a decision scenario where the recommendation given by CDT apparently contradicts our intuitive course of action. Similar to many other alleged counterexamples to CDT, Ahmed's story features an adversary with fantastic predictive power—Death himself, in this story. Unlike many other alleged counterexamples, however, Ahmed explicitly includes the use of a costly randomization device as a possible action for the agent. I critically assess these two features of Ahmed's story. I argue that Death's fantastic predictive power cannot be readily reconciled with the use of randomization device. In order to sustain Dicing with Death as a coherent decision scenario, background explanations must be given about the nature of Death's fantastic predictive power. After considering a few such explanations, however, it becomes unclear if the initial intuition which CDT apparently contradicts still holds up. Finally, I consider two contrasting decision scenarios to illustrate why Ahmed's intuition in this case is ultimately false. I conclude that biting the bullet can perhaps be a legitimate response from CDT to many similar cases where evidentially correlated but causally isolated acts seem to force CDT to give counterintuitive recommendations.

[1] This paper was completed as my undergraduate dissertation in Philosophy at the London School of Economics. I would like to thank Richard Bradley, Miklós Rédei, and Marie Milofsky for their invaluable guidance.

1 Dicing with Death

Suppose you have an unfortunate encounter with Death in Damascus. While he does not claim your life straight away, he reveals to you that your name appears in his appointment book for tomorrow. He says he has a place written for the appointment, and it is either Aleppo or Damascus. If he meets you at the appointed place tomorrow, you die. If he misses you tomorrow, you will survive and live a long life. However, Death is known to be a highly accurate predictor of human actions. While Death's appointments are set long in advance and cannot be changed, it is likely that he would have predicted where you will be at tomorrow. For the avoidance of doubt, further suppose that there are only two places in this world, Aleppo and Damascus. Seeing your conundrum, a cunning merchant approaches and offers you a fair coin. 'For a small fee I will let you know the result of a fair coin toss. You can let the toss decide where you go - Aleppo if it lands heads, or Damascus if it lands tails.' You are interested but remain skeptical. So, you ask the merchant: 'But surely Death has predicted if I will take up your offer?' The merchant replies: 'that's correct, but that is as far as Death's predictive power can go. My coin is truly indeterministic. Even Death cannot predict its outcome.' Staying in Damascus, going to Aleppo, or taking up the merchant's offer? You are left to decide what is the best thing to do.[2]

The unpleasant situation you face is introduced by Arif Ahmed in 'Dicing with Death' (2014). By adding an interesting twist to Gibbard & Harper's classic case Death in Damascus (1978)[3], Ahmed claims to have constructed a decisive counterexample against Causal Decision Theory (CDT)—the decision theory of choice for most philosophers. Since Ahmed's paper was only published relatively recently, there has not been much response from philosophers. Moreover, the few papers I know of that do refer to 'Dicing with Death' mostly take as given Ahmed's conclusion that his case deals a fatal blow to CDT[4]. The aim of this dissertation is to challenge this conclusion. I consider two important features of Ahmed's story, namely Death's fantastic predictive power and randomization as a possible action.

---

[2] Paraphrased from Ahmed 2014 and Gibbard & Harper 1978.

[3] Everything is the same except that there is no merchant with his coin toss offer in the original Death in Damascus case.

[4] See Greene 2017, Gallow 2018, Soares and Levinstein 2017, and Yudkowsky and Soares 2017. All but Greene proposes some alternative decision theory that purports to 'fix' the problem with CDT as exposed by Dicing with Death. Greene discusses the link between rationality and success at a meta-level, which will be relevant in my conclusion.

I argue that Death's fantastic predictive power cannot be taken for granted and some explanations must be given, otherwise the story would be incoherent or at least incomplete. Once we consider such explanations, however, I show that the initial intuition for randomization is misguided. Consequently, Dicing with Death does not present a decisive case against CDT.

The dissertation is organized as follows. In section two, I provide a context for this dissertation by giving an overview of Causal Decision Theory (CDT). In section three, I flesh out Ahmed's argument. In section four, I consider what CDT actually recommends in Dicing with Death and point out the need to articulate the nature of Death's predictive power. In section five, I consider a few plausible explanations. In section six, I illustrate that the initial intuition is misguided by presenting two parallel decision scenarios. I conclude in section seven.

2 CDT in brief

As its name suggests, CDT is a brand of decision theory. Decision theory is concerned with instrumental rationality. It takes an agent's ends as given and considers only how one can best achieve her ends when faced with multiple alternatives. There are three essential ingredients for a decision problem: states of the world, acts available to the agent, and outcomes as specified by act-state pairs. States of the world are conditions of the world that are outside the agent's control (e.g. if it rains today). While the set of all possible states of the world is certainly uncountable, it suffices to provide a finite partition of the universal set in order to define a decision problem[5]. Acts are things that the agent does have direct control over (e.g. whether to bring an umbrella). An outcome is completely determined by an act-state pair and needs to be something that the agent intrinsically cares about. In effect, this ensures that the decision problem is well-formulated by placing a restriction on the partition of the states of the world we use. When we are deciding if we should bring an umbrella out, for example, the partition {spring, summer, autumn, winter} would not be sufficient. It might be the case that it is more likely to rain in some seasons than others, but the agent receives no pleasure nor displeasure by bringing an umbrella in spring per se (assuming the agent does not have aesthetic preferences or the like over the matter). {rainy, not rainy} would be a

[5] This definition of states of the world is slightly different from the standard treatment. Savage (1954), for instance, defines a state as "a description of the world, leaving no relevant aspect undescribed". By requiring that each act-state pair must be of intrinsic value to the agent, I make sure that "no relevant aspect is undescribed".

more appropriate partition, because having an umbrella on a rainy day is something that the agent intrinsically cares about[6].

For each state of the world, *s*, the agent has some belief about how likely s is to transpire. The relative strength of all these beliefs is characterized by a subjective probability function, otherwise known as a credence function. The desirability of each outcome is characterized by a real number, commonly known as utility. Loosely speaking, decision theory calculates the desirability of an act by taking the probability-weighted average of the desirability of all possible outcomes associated with that act. This quantity is called the expected utility of act *a*. Formally,

$$EU(a) = \sum_{s \in S} P(s \text{ if } a) u(a, s)$$

Where *a* is an act, *s* is a state of the world belonging to the partition of possible states of the world S, *P* is the subjective probability function, and *u(a,s)* is a real number that represents the desirability of the outcome as specified by the act-state pair (*a*,*s*). An instrumentally rational agent should choose act(s) with the highest expected utility.

P(s if *a)* is kept vague in the above expression intentionally. Indeed, the appropriate interpretation of this probability is the main contention between Causal Decision Theory (CDT) and Evidential Decision Theory (EDT). For EDT, *P(s if a)* is understood as the usual conditional probability *P(s|a)*. As in mathematics, *P(s|a)* differs from *P(s)* whenever there is correlation between s and a, even if *a* does nothing to causally bring about *s*. For example, while bringing an umbrella does not cause raining, the conditional probability P(rainy|umbrella) is still greater than P(rainy). EDT thus captures the evidential relationship between the agent's acts and states of the world. People often denote EDT expected utility of an act as $V(a)$:

$$V(a) = \sum_{s \in S} P(s \mid a) u(a, s)$$

Intuitively, $V(a)$ measures a's "news value" or auspiciousness. EDT recommends the most auspicious act.

CDT interprets *P(s if a)* differently. A good decision should produce a desirable outcome, rather than merely provide evidence for a desirable outcome. *P(s if a)* should

6 Note that the partition {rainy&spring, rainy&summer, rainy&autumn, rainy&winter, not rainy&spring, not rainy&summer, not rainy&autumn, not rainy&winter}, though unnecessarily fine-grained, is permissible as well, where u(bringing umbrella, rainy & spring) = u(bringing umbrella, rainy) and etc..

capture the causal relationship between *s* and *a*, and ignore any spurious correlation. There are generally two approaches to track causality in the literature. The first approach makes use of counterfactual conditionals. Counterfactual conditionals are propositions of the form "if A were to happen, B would happen". According to this interpretation, *P(s if a)* is understood as the probability of the conditional proposition "If I were to choose *a*, then *s* would happen"[7]. Alternatively, CDT can use the unconditional probability *P(s*) but encode causal information in the choice of the partition of states instead. In the umbrella example, the use of unconditional probability P(rainy) and P(not rainy) is unproblematic because bringing/not bringing an umbrella casually contributes to neither "rainy" nor "not rainy". That is, as long as states are causally independent of acts, *P (s if a)* can be simply replaced by *P(s)* (Savage 1954). David Lewis extends this idea by requiring that the partition is composed of "dependency hypotheses", which are "maximally specific propositions about how states do and do not depend causally on his present action" (Lewis 1981a). In simple terms, dependency hypotheses are an agent's beliefs about the causal structure of the world[8]. The probability associated with each hypothesis is simply the agent's confidence in it. "Bringing an umbrella does not change the weather outside" is a dependency hypothesis. "Having an umbrella in hand creates mini turbulences as you walk. These mini turbulences interact with the atmosphere and eventually causes the weather to turn soggy" is also a dependency hypothesis. For all practical purposes, a rational agent can safely assign probability 1 to the first hypothesis and ignore the second hypothesis. Importantly, while dependency hypotheses specify the extent to which an agent can casually affect states of the world, the hypotheses themselves are causally independent of the agent's actions. Thus, according to Lewis, CDT expected utility can be calculated as[9]:

$$U(a) = \sum_{s \in S'} P(s)u(a, s)$$

[7] It is quite challenging to define the probability of counterfactual conditionals precisely. Doing so requires an in-depth excursion into possible-world semantics (Gibbard and Harper 1978). This is partly why this dissertation opts for the partition approach.

[8] In more recent literature, directed acyclic graphs and a probability distribution over alternative graphs are used to represent dependency hypotheses (Pearl 2009).

[9] Lewis shows that his formulation is equivalent to the counterfactual conditionals approach under background assumptions (Lewis 1981a).

where S' is the set of dependency hypotheses. Intuitively, $U(a)$ measures the causal efficacy of the agent's act. CDT recommends the act that best brings about desirable outcomes.

CDT and EDT gives diverging recommendations when the casual efficacy and news value differ greatly. The contrast is laid bare in Newcomb's problem (Nozick 1969). There are two boxes, box A and box B. You can either take both boxes or box B alone. Box A always has $100 inside. Your adversary decides what is in the box B before you make your decision. He puts $1 million inside if he predicts that you will take only box B, and nothing at all if he predicts that you will take both boxes. He is known to be a highly reliable predictor of your behavior. A sufficiently detailed partition of states of the world is {$1 million in box B, $0 in box B}, denoted as {E, ~E} as a shorthand. Acts available to the agent are {two-box, one-box}. The outcome can be represented in a matrix:

	$1 million in box B	$0 in box B
One-box	1 million	0
Two-box	1 million + 100	100

Without loss of generality, let your adversary's predictive power be quantified as:

P(~E|two-box)=0.99 and P(E|one-box)=0.99

By implication:

P(E|two-box)=0.01 and P(~E|one-box)=0.01

Apply the definition of EDT- expected utility as provided above:

V(one-box)=P(E|one-box)*u(one-box,E)+P(~E|one-box)*u(one-box,~E) =
0.99*1,000,000 + 0.01*0 = 990,000

V(two-box)=P(E|two-box)*u(two-box,E)+P(~E|two-box)*u(two-box,~E)=
0.01*(1,000,000+ 100) + 0.99*100 = 10,100

Clearly, V(one-box) > V(two-box). EDT recommends one-boxing.

In contrast, CDT-expected utility is calculated as:

U(one-box) = P(E)*u(one-box, E)+ (1- P(E))* u(one-box, ~E)

U(two-box) = P(E) * u(two-box, E)+ (1- P(E)) * u(two-box, ~E)

At first glance, it may seem that we are stuck here because the unconditional probability P(E) cannot be readily inferred from the set-up of Newcomb's problem. However, once we plug in values for the desirability of relevant act-state pairs, we see that:

U(one-box) = 1,000,000*P(E)

U(two-box) = (1,000,000 + 100)*P (E) + 100*(1 - P(E)) = 100 + 1,000,000*P (E)

It follows that U(two-box) > U(one-box) for any P(E) between 0 and 1. CDT recommends two-boxing. In some sense, P(E) is only a place-holder. A causal decision theorist does not need to form a rational belief about P(E) in order to settle the decision problem at hand. However, as I will argue later, this will not be the case in Dicing with Death. Perhaps a more intuitive way to understand CDT's position in Newcomb's paradox is the idea of casual dominance[10]. In the state where there is $1 million in box B, two-boxing is preferable to one-boxing. In the state where nothing is in box B, two-boxing is also preferable to one-boxing. Thus, two-boxing is always preferable to box-boxing. Again, this reasoning does not explicitly take into consideration the predictive power of your opponent. What matters is that he made his prediction prior to your decision, and consequently, your decision cannot causally influence what is in the box. Again, we will see that this reasoning breaks down in Dicing with Death.

Most philosophers agree that CDT gets it right in Newcomb's problem[11]. What seems paradoxical is the strong intuition many people feel towards one-boxing when they are first presented with Newcomb's problem. It seems as though by following CDT, people miss out on the opportunity to become millionaires. "If you're so smart, why ain't you rich?" (WAR), this is the standard objection evidential decision theorists lodge against CDT. There is a wealth of responses to WAR from causal decision theorists. Here I briefly explain two arguments that I think will prove relevant to the resolution of Dicing with Death. The first is offered by Joyce (1999), which contends that comparing the fortunes of one-boxers and two-boxers is misleading. WAR asserts that two-boxers would have done better had they chosen one-boxing and one-boxers would have done worse had they chosen two-boxing. However, this is comparing apples to

[10] This is a special case of the Sure Thing Principle with Casual Independence, which states: "For any event E that is causally independent of A and ~A, if the agent prefers A to ~A both on the supposition that E obtains and on the supposition that E does not obtain, then she should prefer A to ~A unconditionally." (Joyce 2007)

[11] In fact, two-boxing is still the right decision even in an enhanced Newcomb's problem where the predictor is infallible (Gibbard and Harper 1978).

oranges. If I have chosen two-boxing, I know that there is most likely nothing in box B; I would have only gotten a worse payoff if I chose one box. If I have chosen one-boxing, I know that there is $1 million in box B; I would still improve my lot by taking both boxes! It is true that two-boxers face worse options (take 100 or get nothing at all) than one-boxers (1 million+100 or 100), but rationality only requires you to optimize given the options you actually have. Gibbard and Harper (1978) give an even simpler explanation to Newcomb's paradox—in a decision situation where the payoff is decided by a good predictor who enjoys rewarding predicted irrationality, then surely irrationality, instead of rationality, is rewarded. Gibbard and Harper's explanation is interesting because it questions the link between rationality and success. I will return to this point in the final part of my dissertation.

Finally, it is worth mentioning that both CDT and EDT belong to subjective, normative decision theory. CDT and EDT are normative because they do not purport to describe how people actually act (descriptive), nor do they try to explain why people act the way they do (explanatory). Instead, they seek to provide decision-making procedures to guide agents' actions, or evaluate the instrumental rationality of agents' actions[12]. CDT and EDT are subjective in the sense that the decision problems are always modelled from the agent's epistemic perspective. It is pointless to require an instrumentally rational agent to have a bird's-eye view in every decision problem she faces. What the agent does not know and cannot find out should not be counted against her. In Dicing with Death, for instance, it should not make any difference to either CDT or EDT's analysis if it turns out that, unbeknownst to the agent, the merchant's offer is in fact a scam. However, this does not mean that an agent can hold any belief whatsoever about her decision situation. If the agent's belief is overly sparse or outright inconsistent, it would not be surprising that even the best decision theory cannot give a good recommendation. This observation will be important to my argument.

3 Ahmed's argument

[12] I am aware that action-guiding and evaluative uses of decision theory are different. An action can be rational according to a decision theory even when the agent did not follow the procedure of that theory, and vice versa (Thoma 2016). For convenience, I am using the term normative to capture both meanings, as the difference between the two will not matter in this dissertation.

Now let's turn back to Dicing with Death and see why Ahmed thinks this case provides a counterexample to CDT. We need some numbers to fill in the details. Suppose you value your life at 10 units and the merchant charges an arbitrary small amount, Δ, for his service. Then the decision problem can be represented in the following matrix:

	S_1: Death in Aleppo & Heads	**S_2: Death in Aleppo & Tails**	**S_3: Death in Damascus & Heads**	**S_4: Death in Damascus & Tails**
Ride to Aleppo	0	0	10	10
Stay in Damascus	10	10	0	0
Coin toss	$-\Delta$	$10-\Delta$	$10-\Delta$	$-\Delta$

(reproduced from Ahmed 2014)

where number in each cell is the utility associated with the corresponding outcome.[13]

[13] Note that acts and states are indeed independent in this table as required by CDT. It is futile to object that the table fails to represent the decision problem faithfully. Consider, for example, the plausible suggestion that if you do not choose to do the coin toss, there would be no "heads" or "tails. So it is false to include the conjunction with "heads" and "tails" in states of the world. A sequential (two-stage) set-up should be used instead. This suggestion is erroneous because even if we did not take up the merchant's offer, the merchant can still flip a coin, thus "generating" these states of the world.

Further, when we have a CDT - permissible partition of the states of the world, we can always expand the set of states of the world by rewriting each state as its conjunction with p and with not p. For example, the Newcomb's paradox, we can harmlessly expand the set of states of the world by interacting each state with the proposition. "the second box is red". We now have four states of the world: {money in box B & box B is red, money in box B & box B is not red, money not in box B & box B is red, money not in box B & box B is not red} yet CDT's recommendation would remain the same.

Formally, let $\mathrm{K} = \{\mathrm{K}_1, \mathrm{K}_2, \ldots, \mathrm{K}_\mathrm{n}\}$ be a CDT-permissible partition of states of the world. Let Q be the set formed by taking the conjunction of each state in K with p and with not p. That is, $\mathrm{Q} = \mathrm{U}_{i=1}^{n}\{k_i \& p, k_i \& \sim p\}$. Note Q will remain CDT-permissible. The CDT-expected utility of an act a with respect to K is

$\mathrm{U_k(a)} = \sum_{i=1}^{n} P(k_i) * u(k_i, a)$.

Let $P(S_1)$ be X_1, $P(S_2)$ be X_2, $P(S_3)$ be X_3, and $P(S_4)$ be X_4. Ahmed proposes to calculate the expected utility of each act according to CDT:

(1) $U(\text{Aleppo}) = 10\ X_3 + 10\ X_4$

(2) $U(\text{Damascus}) = 10\ X_1 + 10\ X_2$

(3) $U(\text{coin toss}) = -\Delta\ X_1 + (10 - \Delta)\ X_2 + (10 - \Delta)\ X_3 + -\Delta\ X_4$

Because this is a fair coin toss, we further know

(4) $X_1 = X_2 = (X_1 + X_2)/2$

(5) $X_3 = X_4 = (X_3 + X_4)/2$

(6) $X_1 + X_3 = 1/2$

(7) $X_2 + X_4 = 1/2$

(4) holds true because given that Death is in Aleppo, Death in Aleppo & H, and Death in Aleppo & T are equally likely. Similarly, (5) holds true.

(6), though not mentioned in Ahmed's paper, holds true because $X_1 + X_3$ is the probability of getting a head, regardless where Death is. Similarly, we have (7).

Use (4) - (7) to replace X_2, X_3, X_4 with X_1 in (1), (2), (3), we have.

(8) $U(\text{Aleppo}) = 10 - 20\ X_1$

(9) $U(\text{Damascus}) = 20\ X_1$

(10) $U(\text{coin toss}) = 5 - \Delta$

In order for $U(\text{coin toss}) > U(\text{Damascus})$ to be true, we must have

$5 - \Delta > 20\ X_1$, which is equivalent to $(5 - \Delta)/20 > X_1$, but this implies

$U(\text{Aleppo}) = 10 - 20\ X_1 > 10 - 20\ (5 - \Delta) / 20 = 5 + \Delta > 5 - \Delta = U(\text{coin toss})$

The expected utility with respect to Q is

$$U_Q(a) = \sum_{j=1}^{2n} P(q_j) * u(q_j, a) = \sum_{i=1}^{n} (P(k_i \& p) * u(k_i \& p, a) + P(k_i \& \sim p) * u(k_i \& \sim p, a))$$

$$= \sum_{i=1}^{n} (P(k_i \& p) + P(k_i \& \sim p)) * u(k_i, a) = \sum_{i=1}^{n} P(k_i) * u(k_i, a) = U_k(a)$$

While CDT is not partition invariant in general (Lewis 1981), this shows that at least Ahmed's set-up in Dicing with Death is not what causes problem for CDT.

It follows that U(coin toss) > U(Damascus) and U(coin toss) > U(Aleppo) cannot both be true. CDT does not recommend coin toss. According to Ahmed, this shows that CDT is false.

I fear that Ahmed is too quick to jump to a conclusion here. As I understand it, Ahmed's argument can be spelt out more fully as follows:

P1: CDT does not recommend taking up the coin toss offer in Dicing with Death

P2: Taking up the offer is the uniquely rational thing to do in Dicing with Death

P3: A decision theory that fails to recommend the uniquely rational thing to do in Dicing with Death cannot be the correct decision theory

Conclusion: CDT cannot be the correct decision theory

I do not dispute P1. Ahmed's mathematical demonstration is correct.

What I have issue with are P2 and P3. Instead of refuting P2 and P3 directly, however, I will first take a closer look at what CDT actually recommends and the assumptions we have to make in order to make sense of the decision situation in Dicing with Death.

4 What does CDT recommend

While CDT does not recommend the coin toss, it is not immediately clear what CDT recommends in Dicing with Death. For convenience, let E be the event that Death is in Damascus and ~E the event that Death is in Aleppo. Recall that (1) U(Aleppo) = 10 X_3 + 10 X_4 and (2) U(Damascus) = 10 X_1+ 10 X_2. Since X_3 + X_4 is precisely the probability that Death is in Damascus, and X_1 + X_2 the probability that Death is in Aleppo, we have

U(Aleppo) = 10 P(E)

U(Damascus) = 10 P(~E) = 10 (1- P(E)) = 10 - 10 P(E).

And also U(coin toss) = 5 – Δ

It follows that U(Aleppo) > U(Damascus) and U(Aleppo) > U(coin toss), whenever P(E) > 1/2, and U(Damascus) > U(Aleppo) and U(Damascus) > U(coin toss) whenever P(E) < 1/2. In other words, going to Aleppo should be chosen if you think that there is more than one half chance that Death awaits you at Damascus, and vice versa. Note this is very different from Newcomb's problem where a causal decision theorist can suspend

judgment about the likelihood of a particular state of the world obtaining and still make a decision. We have to work out what P(E) is. The set-up does not directly tell us what P(E) is, but we can apply Law of Total Probability to relate unconditional probability P(E) to conditional probabilities:

P(E)=P(E|Damascus) * P(Damascus) + P(E|Aleppo) * P(Aleppo) + P(E|coin toss) * P(coin toss).[14]

We are told that Death is an excellent predictor of human behavior, so conditional probabilities P(E|Damascus) should be close to 1, P(E|Aleppo) close to 0. I think it is fair to assign 1/2 to P(E|coin toss), since if Death predicted that you would take the coin toss, the best thing he can do is also randomly pick a place and wait for you there. P(E)≈P(Damascus) + P(coin toss)/2. Substitute into the expression for utilities.

U(Aleppo) ≈ 10P(Damascus)+5P(coin toss)

U(Damascus) ≈ 10P(~E) = 10(1-P(E))=10-10P(Damascus)-5P(coin toss)

Notice that the U (Aleppo) is an increasing function in P(Damascus) and P(coin toss), and, therefore, decreasing in P(Aleppo) = 1- P(Damascus) - P(coin toss). Similarly, U (Damascus) is a decreasing function in P(Damascus). This is the feature of a well-known problem among causal decision theorists – decision instability. Intuitively, the more confident you are that you will go to Aleppo, there is more reason for you to think

[14] Here I assume it makes sense to assign probabilities to acts, at least when the agent is in the process of deliberation. This is a potentially contentious issue among decision theorists. Levi (2000) and Levi (2007) contend that "Deliberation crowds out prediction, so that a decision-maker may not coherently assign unconditional probabilities to the propositions he regards as optional for him". This is commonly known as the DCOP (Deliberation Crowds Out Prediction) thesis. Most recently, Liu, Yang and Price, Huw (2018) attempt to close the debate by saying that criticism of DCOP such as Joyce (2002) is more of a terminological dispute than a genuine debate. The most compelling piece I read on the DCOP debate, however, is Hajek (2016), in which he nullifies top worries about act probabilities and shows that act probabilities are indispensable in deliberation.

A quick note on the interpretation of act probabilities: I take P(A) = p to simply mean that the agent has p degree of belief in the proposition that he will do A at the end of his deliberation. This is a relatively uncontroversial position taken by Harper (1986), and Arntzenius (2008).

Now, even if assigning probabilities to acts is problematic, I take it as a platitude that agents may engage in rudimentary counterfactual thinking ("suppose I choose A, then.."). If I have decided to stay in Damascus, then the probability that Death is in Damascus is close to 1, since Death is an excellent predictor of my behavior. Therefore, the expected payoff of staying in Damascus is close to 0, whereas the expected payoff of going to Aleppo is close to 10. This is sufficient to show that CDT does not recommend either pure option outright.

that Death awaits you there. Consequently, going to Aleppo seems a worse decision to make and you therefore would have less confidence that you will go to Aleppo.

The standard solution to decision instability is deliberational dynamics[15]. The basic idea is simple. The agent starts with some initial assignment of act probabilities, P_0(Damascus), P_0(Aleppo), P_0(coin toss). The only restriction on the initial assignment is that they have to strictly between 0 to 1, since "extreme" values 0 or 1 will make the subsequent updating impossible. With these initial values, the agent evaluates causal utilities for each act, U_0(Damascus), U_0(Aleppo), U_0(coin toss). Now he may find that one of the options, say staying in Damascus, produces higher utility than the other two. But instead of immediately deciding to stay in Damascus, that is, assigning 1 to P(Damascus), he merely updates act probabilities by setting P_1(Damascus)> P_0(Damascus). Of course, he needs to set P_1(Aleppo) and P_1(coin toss) accordingly to make sure P_1(Damascus) + P_1(Aleppo) + P_1(coin toss) = 1[16]. Having updated the set of act probabilities, the agent proceeds to evaluate causal utilities for each act again, U_1(Damascus), U_1(Aleppo), U_1(coin toss). Based on this new set of utilities, the agent again updates act probabilities. The agent continues this procedure until an equilibrium is reached (at the limit), where P_{t+1}(Aleppo)=P_t(Aleppo), P_{t+1}(Damascus)=P_t(Damascus), and P_{t+1}(coin toss)=P_t(coin toss). A unique equilibrium is guaranteed to exist under some mild conditions. While deliberational dynamics is particularly apt in dealing with cases of decision instability, it can be applied to any decision problem. In Newcomb's paradox, for example, the procedure will result in an equilibrium assigning probability 1 to two-boxing and probability 0 to one-boxing.

The only equilibrium in Dicing with Death is P(Damascus)=1/2, P(Aleppo)=1/2, and P(coin toss)=0. Here is a sketch proof. First note that acts with non-zero probability at equilibrium must have equal utility, otherwise the updating procedure would continue by increasing the probability of the act that has higher utility. Suppose at equilibrium we have non-zero probability for coin toss, this would imply that U(Damascus)=U(Aleppo)=U(coin toss). But we have shown in Section 3 that this is impossible. It follows that at equilibrium, P(coin toss)=0 and

[15] The following presentation of deliberational dynamics is a mixture from Joyce (2012) and Arntzenius (2008), although both Joyce and Arntzenius trace their ideas ultimately to Skyrms (1990).

[16] The updating rule is by no means unique, though the agent should always assign higher probabilities for acts that give high utilities and lower probabilities for acts that give low utilities.

U(Damascus)=U(Aleppo) > U(coin toss). This only takes place when P(Damascus) = P(Aleppo)=1/2[17].

What does the outcome of deliberational dynamics say about CDT's recommendation in Dicing with Death? Decision theorists do not have a unanimous answer. Harper (1986) thinks that only the mixed strategy corresponding to the probability distribution at equilibrium is permissible. That is, the agent "will have reasoned himself into becoming a chance device"[18], which decides to stay in Damascus and go to Aleppo with equal probability. Joyce (2012), on the other hand, maintains that once the agent has worked through the deliberational dynamics, any act of non-zero probability or the probability mixture of these acts is permissible. The agent can "pick" an act with an arbitrary tie-breaker. Note this "picking" does not reflect anything about the agent's reasons for doing the act. It is simply meant to prevent an indecision.

If it is indeed possible, as Harper (1986) suggests, for the agent to reason himself into an internal chance device, Dicing with Death poses no threat to CDT at all. The reason for not taking up the offer is simple – why would you pay for randomization when you can do it cost-free? CDT gets Dicing with Death exactly right. I do not see any principled reason why an internal chance device is impossible. Moreover, the chance device does not have to be internal at all. Instead of paying the merchant a penny, for example, the agent might as well flip the penny coin himself.

However, even if we assume that a mixed strategy is unavailable to the agent, the final action that the agent takes will seem stochastic from the agent's epistemic point of view. As Joyce (2012) maintains, whatever the agent ends up doing, he can have "no reason" for doing it. If the agent comes to know that he is going to perform any act, he would have broken the deliberational equilibrium, and, therefore, can no longer be following CDT's recommendation. A question naturally arises—if you yourself cannot know the final outcome of your deliberation, how is it possible that Death has predicted it? This is a serious concern, because as explained in Section 2, for Dicing with Death to count as a decision problem at all, it must be possible for an agent to coherently

[17] Unsurprisingly, this is the same prescription as in the original Death in Damascus (Gibbard and Harper 1978). To borrow a few concepts from game theory, coin toss is strictly dominated by the mixed strategy of 1/2 Damascus + 1/2 Aleppo. Iterated elimination of strictly dominated strategies therefore ensures that having coin toss as an option does not affect what happens at equilibrium.

[18] In fact, Harper continued in the following paragraph: "I assume that Death cannot predict the outcome of the chance device (internal or external) that an agent uses to execute a mixed strategy even though he can predict which mixed strategy gets chosen" (Harper 1986). This is exactly my concern.

believe that he is in such a situation. I am skeptical if this is possible but, for the sake of argument, I consider a few such explanations in the next section.

5 Accounting for predictive prowess

I propose two prima facie plausible explanations that might sustain Dicing with Death: Regression Hypothesis and Simulation Hypothesis. I call these explanations "hypotheses" for a reason. While they are not strictly speaking dependency hypotheses which outline what the agent can and cannot control, they are background beliefs about the underlying causal structure of the Deaths' predictive power. The agent needs these hypotheses to maintain his belief that Death can predict the outcome of his deliberation but not the outcome of the merchant's coin toss.

Regression Hypothesis: "Death has unlimited computational power and detailed information on your past behaviors. In order to make a prediction, he tests a large number of regression models on your past information set and picks the one with the highest degree of fit as his predictive model. Regression is not necessarily frequentist. Death may hold prior on human behavior in general and incorporate the prior information into his predictive models" (RH). RH is plausible because this is what statisticians do in real life for predictive inferences. If your past life indeed follows a very rigid pattern, Death might be able to extract the hidden correlations and make a prediction accordingly. Despite its plausibility, I argue that RH is insufficient to sustain Dicing with Death. As is well-known in statistical finance, history cannot always predict future. Financial models that fit perfectly with historical data perform poorly when there is a fundamental shift in the market structure. In Dicing with Death, even if you visited Aleppo on literally very Monday in your past life, the very fact that you have this unfortunate encounter with Death on Sunday changes where you may go on Monday fundamentally. Whatever predictive model Death uses, it will no longer be indicative of your current deliberation. One might object that Death can model your current encounter with him into his predictive model and make an inference accordingly. However, this is not plausible because Dicing with Death is a singular event. There simply is not any past information that Death can use to model your behavior in Dicing with Death[19]. If you had any encounter with Death, you would have likely perished, whereas if you have survived many Dicing with Death cases, again it would be

[19] This is related to the reference class problem, which is a whole different debate to have.

irrational to believe that Death is good at predicting your behavior. In short, RH fails to account for your belief in Death's amazing predictive power.

One problem with RH seems to be that Death only makes inferences based on outwardly characteristics and does not know how you make a decision. Simulation Hypothesis "fixes" this problem: "Death has unlimited computational power and detailed information on your past behaviors. In addition, he has access to your mental states including all beliefs and desires. Death makes a prediction by simulating your thinking process with exactly the same input information and decision procedure. Death makes a prediction by thinking in your shoes, so to speak" (SH). SH can be a very powerful explanation for the kind of amazing predictive power decision theorists often speak of. In Newcomb's problem, for example, SH can perfectly explain how your adversary manages to predict your choice. By simulating your decision procedure, he sees that you would choose two-boxing and, therefore, predicts that you will choose two-boxing. However, it is not hard to see that SH cannot give Death the predictive power he needs. Death may consider the same partition, draw up the same payoff matrix, and even begin the deliberation with the same initial assignment of act probabilities (for example, set P_0(Aleppo) to 0.99 to incorporate the information that you always visited Aleppo on Mondays). Simulation can take him as far as the unique deliberational equilibrium where none of the pure options is recommended unequivocally, but no more. Since the final action you take will seem stochastic from your epistemic perspective, even a perfect simulation of your deliberation process cannot predict your behavior. SH fails as well.

6 Two Penalty Kick cases

RH and SH certainly do not exhaust all possibilities but I think it is fair to say that they are by far the most intuitive and plausible explanations. This provides strong reason to suspect that Dicing with Death is not a genuine decision problem. Ahmed needs to make questionable assumptions to uphold Dicing with Death as a counterexample to CDT. However, even if we suppose that the belief in Death's predictive power is sustainable and that mixing strategy is unavailable, I maintain that CDT still gets Dicing with Death right. Death has made his prediction and he either seeks you in Damascus or Aleppo. In either case, your decision can do nothing to his whereabouts tomorrow. You are just as likely to encounter Death wherever you end up going. I submit that this is highly counterintuitive. Many people would feel strongly that taking up the offer is the right thing to do. However, I think this intuition is misguided, precisely because of the ambiguity in the nature of Death's predictive power. I fear that when people read Dicing with Death, they implicitly assume that they are in

one decision situation when they are really in another. Consider the following two cases for an illustration.

Penalty Kick 1: “You are tasked to execute a penalty kick. Unfortunately, your opponent team is known to have a formidable goalkeeper. His career record shows that he saved 99% of penalties he ever faced. You know that he pays close attention to your micro-muscle moves, facial expressions and so on, in order to make a prediction on which side you will kick (left or right for simplicity). Before you go on to the pitch, however, your coach gives you a little gadget that emits an electric current once you press the button on it. The electric shock can trigger a special neural pathway in your brain and instantaneously randomize the direction in which you place your penalty kick, without affecting your power and precision. As a side effect, the electric shock will leave a small scar in your hand but is otherwise harmless. Should you use the gadget? And when?”

Most people will agree that using the gadget at the precise moment when you kick the ball is the uniquely rational decision in Penalty Kick 1. If you kick either left or right, you are almost certain that the goalkeeper will bask in glory, as he always did, at the expense of your misery. If you use the gadget, however, you can score with probability one half. Indeed, CDT recommends using the gadget as well. There is no need to go through the whole formal apparatus to verify CDT’s recommendation. The crucial fact in Penalty Kick 1 is this – the goalkeeper has not made his prediction. By using the gadget, you effectively nullify the goalkeeper’s predictive power by making the information based on which he makes his prediction irrelevant. Using the gadget causally promotes the desirable outcome of scoring.

Now consider a different case. Penalty Kick 2: “As in Penalty Kick 1, you are tasked to execute a penalty kick. The goalkeeper you face is equally formidable with 99% success rate in penalty saves. Except this time, you know he has a very peculiar habit. Because he is so confident of his skills, he makes a prediction about which side you kick the moment you step onto the pitch, and closes his eyes before making the save. You still have the gadget your coach gives in Penalty Kick 1. Should you use it?”

I hope most people share my intuition that one should not use the gadget in Penalty Kick 2. The goalkeeper has made his decision on which direction he will save. Using the gadget will get you nothing but a scar. I suspect that people feel strong intuition for using the coin toss in Dicing with Death because they implicitly assumed they are in Penalty Kick 1, when Dicing with Death is really more like Penalty Kick 2. Death’s predictive power is a mystery, but the natural inclination is to believe that prediction requires causal information. By using the coin toss, people assumed that you can

somehow invalidate the information on which Death makes his prediction as in Penalty Kick 1. However, this cannot be the case. Plainly, Death does not causally depend on information about your decision to make a prediction. This is also why people have strong intuition for one-boxing when they face an infallible predictor in Newcomb's paradox, even though two-boxing is, upon reflection, the only rational option. To account for an infallible predictor, people subconsciously discounted their conviction that backward causality is impossible. When people feel strongly for one-boxing, they are simply motivated by the implicit belief that current choice may affect what is in the box, which we know is impossible (McKay 2004). The strong intuition for taking up the offer in Dicing with Death is motivated by precisely the same psychology.

Let no one object that Penalty Kick 2 is impossible because no goalkeeper can close his eyes and still make the save. This is exactly my concern with Dicing with Death. Death made his prediction when he wrote down your name in the appointment book, just as the goalkeeper made his prediction as soon you stepped on the pitch. When Death meets you this morning, he has already closed his eyes yet you have no trouble believing that he knows where you will be tomorrow. If we are to accept Death's amazing predictive power, we should not have a problem with the goalkeeper's skill. Taking up the merchant' offer is not rational.

7 Conclusion

Finally, a word about rationality and success. I submit that causal decision theorists will fare badly in this situation—provided that we assume that Dicing with Dicing is a genuine decision problem and that mixed strategy is unavailable. However, CDT performs badly in Dicing with Death not because the theory itself is dysfunctional, but rather because of the bizarre decision situation and the no-mixed strategy restriction Ahmed secretly placed on the agent. Taken to the extreme, one can easily conjure up a demon which has perfect information about you and punishes you if you deliberate as a causal decision theorist. This can hardly be a decisive objection to CDT. As in Newcomb's problem, rationality only requires you to optimize given the cards you are dealt with. Your encounter with Death deals you a bad hand and CDT should not be blamed for not being able to help you any better.

Bibliography

Ahmed, Arif. 2013. “Causal Decision Theory: A Counterexample.” *Philosophical Review* 122 (2): 289–306. https://doi.org/10.1215/00318108-1963725.

———. 2014. “Dicing with Death.” *Analysis* 74 (4): 587–92. https://doi.org/10.1093/analys/anu084.

Arntzenius, Frank. 2008. “No Regrets, or: Edith Piaf Revamps Decision Theory.” *Erkenntnis (1975-)* 68 (2): 277–97.

Bales, Adam. 2018. “Richness and Rationality: Causal Decision Theory and the WAR Argument.” *Synthese* 195 (1): 259–67. https://doi.org/10.1007/s11229-016-1214-x.

Egan, Andy. 2007. “Some Counterexamples to Causal Decision Theory.” *Philosophical Review* 116 (1): 93–114. https://doi.org/10.1215/00318108-2006-023.

Gibbard, Allan, and William L. Harper. 1978. “Counterfactuals and Two Kinds of Expected Utility.” In *IFS*, 153–90. The University of Western Ontario Series in Philosophy of Science. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-9117-0_8.

Greene, Preston. 2017. “Success-First Decision Theories,” Forthcoming in Newcomb's Problem, Cambridge University Press.

Hájek, Alan. 2016. “DELIBERATION WELCOMES PREDICTION.” *Episteme* 13 (4): 507–28. https://doi.org/10.1017/epi.2016.27.

Harper, William. 1986. “Mixed Strategies and Ratifiability in Causal Decision Theory.” *Erkenntnis (1975-)* 24 (1): 25–36.

Joyce, James M. 1999. “The Foundations of Causal Decision Theory.” Cambridge Core. April 1999. https://doi.org/10.1017/CBO9780511498497.

———. 2007. “Are Newcomb Problems Really Decisions?” *Synthese* 156 (3): 537–62. https://doi.org/10.1007/s11229-006-9137-6.

———. 2012. “Regret and Instability in Causal Decision Theory.” *Synthese* 187 (1): 123–45.

———. 2016. Book Review “Arif Ahmed: Evidence, Decision and Causality.” *The Journal of Philosophy* 113 (4): 224–232. https://doi.org/10.5840/jphil2016113413.

Lewis, David. 1981a. "`Why Ain'cha Rich?'." *Noûs* 15 (3): 377–80. https://doi.org/10.2307/2215439.

———. 1981b. "Causal Decision Theory." *Australasian Journal of Philosophy* 59 (1): 5–30. https://doi.org/10.1080/00048408112340011.

Liu, Yang, and Huw Price. 2018. "Ramsey and Joyce Deliberation and Prediction." Preprint. March 3, 2018. http://philsci-archive.pitt.edu/14426/.

McKay, Phyllis. 2004. "Newcomb's Problem: The Causalists Get Rich." *Analysis* 64 (2): 187–89.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel*, 114–46. Synthese Library. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-1466-2_7.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2016.

Savage, Leonard J. Wiley and Sons. *The Foundation of Statistics*, 1954, pp. 6–21.

Skyrms, Brian. *The Dynamics of Rational Deliberation*. Harvard Univ. Press, 1990.

Soares, Nate, and Benjamin A Levinstein. 2017. "Cheating Death in Damascus," *Formal Epistemology Workshop (FEW) 2017* University of Washington, Seattle, USA. May 26–28, 2017.

Weirich, Paul. 1985. "Decision Instability." *Australasian Journal of Philosophy* 63 (4): 465–472.

Yudkowsky, Eliezer, and Nate Soares. 2017. "Functional Decision Theory: A New Theory of Instrumental Rationality." *ArXiv:1710.05060 [Cs]*, October. http://arxiv.org/abs/1710.05060.