Cheating Death in Damascus

Benjamin A. Levinstein
University of Illinois at Urbana-Champaign
benlevin@illinois.edu

Nate Soares
Machine Intelligence Research Institute
nate@intelligence.org

Abstract

Evidential and Causal Decision Theory are the leading contenders as theories of rational action, but both face fatal counterexamples. We present some new counterexamples, including one in which the optimal action is causally dominated. We also present a novel decision theory, Functional Decision Theory (FDT), which simultaneously solves both sets of counterexamples. Instead of considering which physical action of theirs would give rise to the best outcomes, FDT agents consider which output of their decision function would give rise to the best outcome. This theory relies on a notion of subjunctive dependence, where multiple implementations of the same mathematical function are considered (even counterfactually) to have identical results for logical rather than causal reasons. Taking these subjunctive dependencies into account allows FDT agents to outperform CDT and EDT agents in, e.g., the presence of accurate predictors. While not necessary for considering classic decision theory problems, we note that a full specification of FDT will require a non-trivial theory of logical counterfactuals and algorithmic similarity.

1 Introduction

Evidential and Causal Decision Theory (EDT and CDT) are the two main contenders as theories of rational action, but neither has managed to persuade enough critics to establish itself as the consensus choice. This division is in part due to the fact that philosophers disagree about the correct solution to several decision problems, such as Newcomb's problem. In order to resolve this dispute, we will explore some problems that we believe should be uncontroversially fatal to both theories. The flaws of EDT and CDT will then lead us to a new theory called *functional decision theory* (FDT) that not only gets these less controversial problems right but has strong theoretical motivation as well.

The decision problems that we focus on are akin to Gibbard and Harper's (1978) classic Death in Damascus problem. Traditional solutions to such problems are unstable. Once you grow confident you'll perform one action, you will expect that action to produce bad consequences. These cases lead to more univocal intuitive verdicts, so they provide probative data for deciding between theories.

EDT gets these cases right, while CDT gets them wrong. Nonetheless, EDT faces its own well-known and (to our minds) decisive counterexamples. Although it arrives at the right answer in the unstable problems we'll discuss, it does so for the wrong reason.

In short, EDT tells you to perform the action that would be the best *indication* of a good outcome, whereas CDT tells you to perform the action that will tend to *causally bring about* good outcomes. Unfortunately, the former treats spurious correlations

Research supported by the Machine Intelligence Research Institute (intelligence.org). Forthcoming in *The Journal of Philosophy* (journalofphilosophy.org).

as decision-relevant; the latter treats decision-relevant correlations as spurious. FDT is structurally similar to CDT, but it rectifies this mistake by recognizing that logical dependencies are decision-relevant as well.

To get the gist of the differences between the three theories, consider the following problem as illustration:

Twin Prisoner's Dilemma An agent has the option of cooperating with or defecting against her psychological twin, who is in a separate room, faces the same options, and has the same state of knowledge. Both rank the outcomes the same way, where "I defect and she cooperates" > "We both cooperate" > "We both defect" > "I cooperate and she defects." Should the agent defect or cooperate?

CDT reasons that because there's no causal link between the two agents' choices, defecting strictly dominates cooperating. So, it prescribes defection. EDT reasons that defection is strongly correlated with the other agent defecting and that cooperation is correlated with the other agent cooperating. Since it prefers the mutual cooperation outcome to the mutual defection outcome it prescribes cooperation.

The FDT agent reasons as follows: *If my decision algorithm were to output* defect*, then the other agent would output* defect *too because she's running (a close approximation of) my algorithm. If my algorithm were to output* cooperate*, then the other agent would choose that as well. The latter results in a better outcome for me, so I cooperate.*

More generally, functional decision theorists think of agents as instantiations of a particular decision algorithm (combined with credences and utilities). FDT advises agents to consider what would happen if their decision algorithm were to produce a different output and to choose the output that tends toward the best results. Because functions can be multiply instantiated, all instantiations of that function will counterfactually co-vary (according to the FDT-counterfactuals in the hypothetical scenarios imagined by an FDT agent).

According to CDT, you should only take potential effects that depend *causally* on your action into account, where causation is restricted to the physical dynamics of the universe. Functional decision theory has a different notion of *dependence*. The output of your own algorithm and the output of your twin's algorithm are subjunctively interdependent. You have control over your algorithm (since you are free either to cooperate or defect) and therefore you also have control over the algorithm that your twin implements (because those are the same algorithm). Thus, the correlation between your action and the twin's is not spurious, and your counterfactual reasoning should take this connection between your action and the twin's action into account.

We claim such reasoning is both natural and leads to a better theory than either CDT or EDT. FDT gets the right answer in unstable problems like DEATH IN DAMASCUS while avoiding the pitfalls of EDT.

Functional decision theory has been developed in many parts through (largely unpublished) dialogue between a number of collaborators. FDT is a generalization of Dai's (2009) "updateless decision theory" and a successor to the "timeless decision theory" of Yudkowsky (2010). Related ideas have also been proposed in the past by Spohn (2012), Meacham (2010), Gauthier (1994), and others.

FDT does face new conceptual problems of its own. Most importantly, on any formulation we're aware of, a full specification of FDT requires a notion of non-trivial logical counterfactuals (with logically impossible antecedents) and a notion of similarity between algorithms. We will leave exploration of these topics to future work, since the machinations and verdicts of FDT will be clear in the cases under consideration.

Here's the plan. Section 2 gives a whirlwind summary of EDT and explains why it fails. Section 3 describes CDT, discusses Joyce's (2012) extension of CDT to cover unstable cases like DEATH IN DAMASCUS, and presents Ahmed's (2014) counterexample. We argue that this counterexample shows that CDT's dependency hypotheses are systematically impoverished. This oversight motivates FDT, which

we develop in section 4. Section 5 runs through a number of cases that contrast CDT and FDT. Our first case, DEATH ON OLYMPUS, presents an option which FDT rightly recognizes as optimal despite being causally dominated. PSYCHOPATH BUTTON characterizes FDT further. Our final case, MURDER LESION, is an alleged counterexample to FDT. However, we argue that when correctly spelled out, FDT arrives at the right verdict. Section 6 wraps up.

2 Evidential Decision Theory

We begin with a familiar rehearsal of the motivations and machinations behind the two current leading theories of ideal rational action.

According to Evidential Decision Theory, you should perform the action that is the best indicator of good outcomes. That is, EDT tells agents to perform the act that leads to the greatest expected value conditional on performing it.

To spell this view out abstractly, we'll loosely follow Savage's (1972) model of decision making: an agent uses her credences about which *state of the world* is actual to choose between possible *actions* that lead to better or worse *outcomes*. With Jeffrey (1983), we'll think of both states and actions as propositions: elements of the set $\mathcal{S} = \{s_1, \ldots, s_n\}$ and $\mathcal{A} = \{a_1, \ldots, a_m\}$ respectively. Jointly, an action a and a state s determine an outcome $o[a, s] \in \mathcal{O}$. Outcomes are the objects of intrinsic value, or final ends for the agent. So, $o_1 = o_2$ only if the agent is indifferent between o_1 and o_2.

We take for granted that the agent comes equipped with a (probabilistically coherent) credence function P that measures her subjective degrees of confidence and a utility function $u : \mathcal{O} \to \mathbb{R}$ that represents how good or bad she judges each outcome.

According to EDT, agents should perform the action that maximizes

$$U_{\text{EDT}}(a) = \sum_{s \in \mathcal{S}} P(s \mid a) u\left(o[a, s]\right) \tag{1}$$

That is, EDT agents calculate the expected utility of actions by considering their likely consequences on the *indicative supposition* that the action is performed.

The problem, however, is that equation (1) requires agents to take into account every correlation between their actions and the state of the world, even if those correlations are spurious. As Lewis (1981) puts it, EDT endorses "an irrational policy of managing the news" (p. 5). To help illustrate this flaw, consider:[1]

XOR Blackmail An agent has been alerted to a rumor that her house has a terrible termite infestation, which would cost her $1,000,000 in damages. She does not know whether this rumor is true. A greedy and accurate predictor with a strong reputation for honesty has learned whether or not it's true, and drafts a letter:

> I know whether or not you have termites, and I have sent you this letter iff exactly one of the following is true: (i) the rumor is false, and you are going to pay me $1,000 upon receiving this letter; or (ii) the rumor is true, and you will not pay me upon receiving this letter.

The predictor then predicts what the agent would do upon receiving the letter, and sends the agent the letter iff exactly one of (i) or (ii) is true. Thus, the claim made by the letter is true. Assume the agent receives the letter. Should she pay up?

1. We use this case due to Soares and Fallenstein (2015) instead of the more familiar Smoking Lesion from Skyrms (1980) for two reasons. First, the smoking lesion problem requires the agent to be uncertain about their own desires. Second, the present case is immune to the so-called "tickle defense" (Eells 1984).

What she does now will not affect whether she has termites or not and will in no way improve her situation. While paying the blackmail is a good indication that there are not termites, this correlation between her action and the state of the world is *spurious*. How frequently people end up with termite infestations does not change regardless of their willingness to pay blackmailers of this kind. (Some readers may feel a temptation to pay the blackmailer anyway, which we hope will fade after reading FDT's approach to the problem in Section 4.3.)

3 Causal Decision Theory

Although EDT still has a number of adherents, we think cases like XOR Blackmail show it is fatally flawed. Rational action requires trying to bring about good outcomes. What matters is how the world *depends upon* your actions. Instead of asking what is likely the case given that you perform a, you should ask how the world would change *were* you to perform a.

Accordingly, Causal Decision Theory does not think every correlation between act and state is relevant to decision making, unlike EDT. Instead, it directs agents to consider only the potential causal consequences of their actions. Not paying the blackmail is strong evidence that your house is infested, but paying would not cause a better outcome because paying the blackmailer will not actually bring about a lower chance of an infestation. In fact, paying is sure to cause a worse outcome regardless of whether or not there's an infestation. Thus, you should not pay.

One natural way to spell out CDT is to invoke the notion of *dependency hypotheses* (Joyce 1999; Lewis 1981). A dependency hypothesis is a conjunction of counterfactuals of the form $\bigwedge_{a \in \mathcal{A}} a \mathbin{\square\!\!\rightarrow} s$, where s is a state of the world. I.e., a dependency hypothesis takes a stand on which state of the world would result from any action. The counterfactual conditional $\mathbin{\square\!\!\rightarrow}$ is interpreted causally: if I *were* to perform a, s *would* occur.

CDT Agents form credences over competing dependency hypotheses, which they use to guide their decisions. $P(a \mathbin{\square\!\!\rightarrow} s)$ is greater than, equal to, or less than $P(s)$ exactly in the cases where the agent judges a to causally promote, be causally independent of, or causally inhibit bringing about state s. According to CDT, these counterfactuals encode how the state of the world (and the outcome) are seen by the agent to depend on her actions.

In turn, we can define causal expected utility as:

$$U_{\text{CDT}_1}(a) = \sum_{s \in \mathcal{S}} P(a \mathbin{\square\!\!\rightarrow} s) u\left(o[a, s]\right)$$

Given a probability function that is defined over a set of dependency hypotheses, CDT tells agents to maximize causal expected utility. In contrast to EDT, this equation weights the utility of an outcome (a state-action pair) by the probability of the state on the *subjunctive* supposition that the action is performed.

While there are a number of different versions of CDT and ways of understanding $\mathbin{\square\!\!\rightarrow}$, these differences will not matter to us. For expositional purposes, however, it will be useful to identify $P(a \mathbin{\square\!\!\rightarrow} s)$ with $P(s \mid \texttt{do}(\textsc{Act} = a))$ in accord with Judea Pearl's (1996; 2009) `do`-calculus. In this version, an agent is equipped with both a probability function P and a directed graph G. The nodes of G represent variables potentially relevant to the outcome, and the edges describe potential directions of causal influence.[2]

$P(\cdot \mid \texttt{do}(\textsc{Var} = \mathit{val}))$ is P modified so that all variables causally downstream (according to G) of Var, including Var itself, are updated to reflect the intervention setting Var equal to *val*, and the other variables are left unchanged. To calculate expected utility, CDT agents intervene on the Act variable and use $P(\cdot \mid \texttt{do}(\textsc{Act} = a))$ to calculate the expected value of an action a. So, our official equation for CDT's

2. More generally: a probability distribution over possible graphs. See *fn.* 3.

expected utility formula is:[3]

$$U_{\text{CDT}}(a) = \sum_{s \in \mathcal{S}} P(s \mid \texttt{do}(\text{ACT} = a))u\left(o[a, s]\right) \tag{2}$$

For example, in figure 1, whether there's an infestation affects whether the predictor sends the letter, which also affects the agent's decision to pay. Together, the agent's decision and whether there's an infestation determine her payoff. To calculate the value of paying, the agent uses $P(\cdot \mid \texttt{do}(\text{ACT} = \text{pay}))$. Since the Predictor variable is upstream of ACT, $P(\text{infestation} \mid \texttt{do}(\text{ACT} = \neg\text{pay})) = P(\text{infestation} \mid \texttt{do}(\text{ACT} = \text{pay})) = P(\text{infestation})$, even though $P(\text{infestation} \mid \text{pay}) < P(\text{infestation} \mid \neg\text{pay})$. This makes sense. If the agent were to refuse to pay, she would not change whether there's an infestation or not. The correlation between not paying and the lack of an infestation is thereby counterfactually broken and correctly diagnosed as spurious.

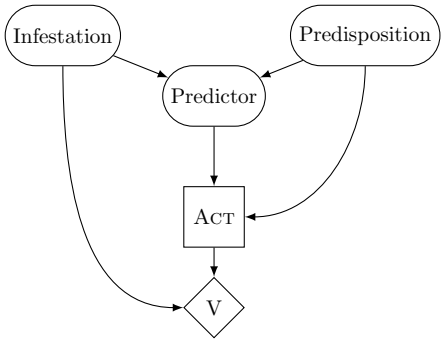


Figure 1: A causal graph for CDT agents facing XOR BLACKMAIL. The agent intervenes on the ACT variable, setting it either to *pay* or *do not pay*. *Infestation* is either 'yes' or 'no', and *Predictor* is either *send* or *do not send* depending on whether there's an infestation and on the agent's predisposition. Whether there's an infestation and whether ACT is *pay* or not determines the agent's payoffs, represented by the V-node.

3.1 CDT and Instability

To complete the discussion of Causal Decision Theory, we'll now consider a new kind of case that appears unstable for CDT. First, we'll discuss how one leading version of CDT might handle such cases and argue that it is doomed to get the wrong answer. We'll then use the reasons for this failure—namely, that CDT has an impoverished conception of dependence—as motivation for FDT. After presenting FDT, we provide a more general counterexample (DEATH ON OLYMPUS) to any standard version of CDT.

We'll begin with an asymmetric variant of a problem originally discussed by Gibbard and Harper (1978):

Death in Damascus You are currently in Damascus. DEATH knocks on your door and tells you I AM COMING FOR YOU TOMORROW. You value your life at

3. Equation (2) is not sufficiently general because the agent may be uncertain which causal graph represents her decision problem. So, CDT in its general form requires credences over not just states and actions, but also over graphs. We will use P as the joint distribution over all three, which we marginalize out as necessary. Therefore, the generalized equation for our version of CDT is:

$$U_{\text{CDT}}(a) = \sum_{i} \sum_{s \in \mathcal{S}} P(G_i) P(s \mid \texttt{do}_i(\text{ACT} = a))u\left(o[a, s]\right)$$

where G_i is a graph, and $\texttt{do}_i$ is the $\texttt{do}$-operator for G_i. Because the structure of the graph is known for all cases in this paper, we avoid this added complexity in the main presentation.

$1,000 and would like to escape DEATH. You have the option of staying in Damascus or paying $1 to flee to Aleppo. If you and DEATH are in the same city tomorrow, you die. Otherwise, you will survive. Although DEATH tells you today that you will meet tomorrow, he made his prediction of whether you'll stay or flee yesterday and must stick to his prediction no matter what. Unfortunately for you, DEATH is a perfect predictor of your actions. All of this information is known to you.

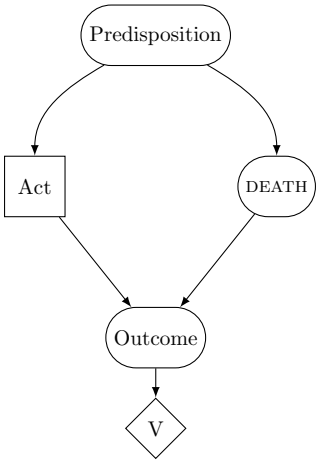


Figure 2: A causal graph for CDT agents facing DEATH IN DAMASCUS, RANDOM COIN, and DEATH ON OLYMPUS. In the first two cases, the agent can set ACT to either *stay* or *flee* and has the additional option of *randomize* in the second case. In DEATH ON OLYMPUS, the agent can stay in Damascus, flee to Aleppo, or *climb* Olympus. The DEATH variable is either *Damascus* or *Aleppo* in the first two cases, and possibly *Olympus* in the third.

First, note that because DEATH has privately made his decision of where to go before speaking to you, your decision of whether to stay in Damascus or flee to Aleppo is causally independent of DEATH's decision. That is, DEATH's decision of where to go does not causally affect your decision, and your decision does not causally affect DEATH's decision.

However, your decision is strong evidence for what the causal structure of the world is. Recall, CDT requires you to consider credences about what *would* happen if you were to perform an action, i.e., credences of the form $P(s \mid \texttt{do}(a))$. In the cases we've discussed, $P(a)$ does not affect the value of $P(s \mid \texttt{do}(a))$. In this case, it does, because DEATH is a perfect predictor.[4]

More explicitly, let S and F refer to the actions of staying in Damascus and fleeing to Aleppo, and let D and A denote the propositions that DEATH is in Damascus and that DEATH is in Aleppo. Suppose you're sure you'll stay in Damascus. That is, $P(S) = 1$. You're then sure that that DEATH will be in Damascus tomorrow. But, because your choice is causally independent of DEATH's location (i.e., DEATH's location is neither upstream nor downstream of your choice on your causal graph), $P(D \mid \texttt{do}(S)) = P(D \mid \texttt{do}(F)) = P(D) = 1$. In this case, fleeing to Aleppo brings you more causal expected utility because you think that fleeing to Aleppo would cause you to live. However, if you're sure you'll go to Aleppo, then your opinions about the causal consequences of your actions are different. In that circumstance, staying in Damascus brings you more causal expected utility.

As we see, the problem in scenarios like DEATH IN DAMASCUS is that the agent cannot be certain of what she'll do before she decides, on pain of irrationality. If she's

4. Some readers may object to the stipulation that DEATH is a *perfect* predictor. This is inessential and merely simplifies the discussion. For our purposes, little is lost if DEATH is a nearly perfect predictor instead.

sure she'll go to Aleppo, then either she as a matter of fact remains in Damascus despite her certainty, or she performs an action that violates the injunction to maximize expected causal utility.[5]

Now, there is a bit of disagreement about what CDT recommends in cases such as these. One option is to say that CDT fails to render a verdict at all. This route is certainly unattractive, as it prevents CDT from being a general theory of rational action, so we won't consider it further.

Recently, Arntzenius (2008) and Joyce (2012) have presented solutions, both of which rely primarily on the model of rational deliberation developed in Skyrms (1990). The differences between Arntzenius's and Joyce's views will not matter much for us, so we'll primarily focus on Joyce's view.

According to Joyce, because the agent cannot know *ex ante* what she'll do, she must initially have non-extremal credences over her actions. To make matters concrete, let P_0 be her initial credence function, and suppose $P_0(S) = .9$ and $P_0(F) = .1$. If she uses this credence function to make her decision, she assigns fleeing higher expected utility than remaining. To be precise: $U_0(S) = 100$ and $U_0(F) = 899$.

It's tempting here to conclude that CDT requires her to flee, but that verdict is premature. Although she initially evaluates F as a more favorable action than S, this very evaluation is important information for a number of reasons:

1. The agent regards herself as rational, i.e., as a causal expected utility maximizer. However P_0 assigns lower credence to F than to S even though she currently evaluates F as a better option than S. Surely, this should lead her to *raise* her credence in F and lower it in S.

2. Given (1), the fact that she regards F as more attractive than S is evidence about DEATH's location. Once she raises her credence that she will go to Aleppo, she should become more confident that DEATH is in Aleppo as well.

So, $P_0(\cdot \mid U_0) \neq P_0$.[6] Thus, P_0 did not take into account all freely available information—*viz.*, the value of U_0. So, the agent should not use P_0 to guide her final action. Instead, she should revise her credence to $P_1(\cdot) = P_0(\cdot \mid U_0)$ and re-evaluate her options. Of course, it might be that $P_1(\cdot \mid U_1) \neq P_1$, in which case she should iterate this process.

It's a subtle matter how exactly she should update on her own utility calculations, but the important requirements are just that she grows more confident she'll perform attractive options, and that she does not instantly become certain she'll perform such options.

Eventually, if she updates in the right way, this iterative process will terminate in an equilibrium state P_e, such that $P_e = P_{e+1} = P_e(\cdot \mid U_e)$. In this case, equilibrium occurs when the agent is *indifferent* between fleeing and remaining, where:

- $P_e(S) = .5005$
- $P_e(F) = .4995$

At this point, the agent has taken into account all freely available, causally relevant information, so P_e is her credence function at the end of deliberation. Strictly speaking P_e only tells her what she *believes* she'll do at the end of deliberation. There remains the further question of what she actually should *do*.

CDTers disagree about this question, and Joyce says that she is permitted either to go to Aleppo or stay in Damascus. Arntzenius thinks that CDT only ultimately tells you about what you should believe and is silent on which *actions* are rational. This latter position strikes us as unattractive, since we want decision theory to guide our actions, not just to tell us what we should think.

5. The former option is at the very least uncomfortable for the CDT theorist. Decision theory should lead to verdicts about what to do at the end of rational deliberation. If the verdict is different from what the agent actually should do, then that is the fault of the underlying decision theory.

6. We use the notation $P_0(\cdot \mid U_0)$ to refer to P_0 conditioned on the value of U_0.

For the sake of generality, we'll now assume the agent has not only the pure actions of staying in Damascus and fleeing to Aleppo, but also has a certain type of *mixed* action available to her as well. She can, if she wishes, flip a mental coin (or use some randomization process) of any bias and let the outcome of that coin flip determine what she ends up doing. Given such an expanded action space, she'll most want to flip a coin with biases matching the equilibrium credences over the act space.[7] So, in this case, she will flip a coin that she has credence .5005 will land heads and will stay in Damascus only when it does.

However, in keeping with the set-up of the case and unlike in standard treatment of mixed acts in decision-theory, we assume any mental randomization she performs is still *predictable from* DEATH*'s perspective*. If it weren't, then DEATH could not be a perfect or nearly perfect predictor, which is the entire motivation for the original case. The coin, then, is not truly random, but has no detectable pattern from the agent's perspective beyond the frequency with which it lands heads or tails. Moreover, the agent is aware that the outcome of the coinflip is predictable to DEATH. We'll call such a coin *pseudo-random*.

So, CDT agents remain in Damascus just over half the time, and they flee to Aleppo just under half the time. In each case, they end up dead.[8]

3.2 Random Coin

Although Joyce's solution in this initial case seems natural, it does lead to an odd verdict. You will sometimes end up paying money to flee to Aleppo *even though you're sure to die there*. It's true, if we use CDT counterfactuals, that when you stay in Damascus, you would have lived if you had fled. And it's true that when you flee, you would have lived if you had stayed. However, as a matter of fact you will always die, and you know that you will always die. The CDT counterfactual provides cold comfort regardless of where you end up. Why pay $1 when you do not have to?

We can bring out the central problem with this solution with a case due to Ahmed (2014):

Random Coin As before, DEATH tells you I AM COMING FOR YOU TOMORROW, and you die only if you end up in the same city. As before, DEATH has made his prediction long ago, and he cannot change it at this point. You now have the option of staying in Damascus or fleeing (for free) to Aleppo. This time, however, you run into Rhinehart, who offers to sell you a truly indeterministic coin for $1 that lands heads with probability .5. Although DEATH can predict whether you will buy the coin, he cannot do better than chance at predicting how the coin lands. On the other hand, if you do not buy the coin, DEATH is a perfect predictor of where you'll end up.

It seems clear that you really should buy the coin. If you do not, you will die with certainty. If you do, you have a 50% chance of survival. By making yourself unpredictable to DEATH *even after* DEATH *has made his irrevocable prediction*, you end up better off.

However, CDT—at least on the versions currently under discussion—advises you not to pay. To see why, note that, structurally, the causal graph here is the same as in Figure 2. The only difference is that in this variant, you have a new potential action: Pay $1 for a truly random coin.

Buying the coin is causally independent of DEATH's location. If DEATH is in Aleppo, the coin will not affect that, and the same goes for Damascus. With or without the coin, you are free to go to either place, and where you go does not

7. To see why, imagine she flipped a coin she has credence .6 will land heads and decides to stay in Damascus when and only when it does. She'd then have credence .6 she'd end up in Damascus instead of retaining her equilibrium credence of .5005. So, after adopting credence .6 in staying, she'd view staying as more attractive.

8. If we restrict our attention to the case in which the agent can't perform mixed acts, then CDT will still allow agents to pay to go to Aleppo. Our case DEATH ON OLYMPUS below will serve as a counterexample to CDT regardless of whether the action space is pure or mixed.

causally affect where DEATH is. So, buying the coin costs a dollar, and it does not causally bring about a better world.

More explicitly, we can calculate the value of your options of staying and fleeing (S and F) as follows:

$$U(S) = P(D \mid \texttt{do}(S))u(S, D) + P(A \mid \texttt{do}(S))u(S, A)$$
$$U(F) = P(D \mid \texttt{do}(F))u(F, D) + P(A \mid \texttt{do}(F))u(F, A)$$

Again, D (A) refers to DEATH ending up in Damascus (Aleppo). Whether you're in equilibrium or not, one of these options always has expected utility of at least 500, since staying and fleeing themselves cost nothing.

Randomizing has a 50% chance of resulting in fleeing and a 50% chance of resulting in staying, but is sure to cost \$1. So,

$$\begin{aligned} U(R) = {} & P(D \mid \texttt{do}(R))[.5(u(S, D) + u(F, D)) - 1] \\ & + P(A \mid \texttt{do}(R))[.5(u(S, A) + u(F, A)) - 1] \\ = {} & 499 \end{aligned}$$

Something must have gone wrong here. Regardless of any theoretical pre-commitments, we have a hard time believing even trenchant defenders of CDT would not buy the coin if they found themselves in this situation.[9]

And the reason to buy a coin seems deeper than mere correlation. Although DEATH already made his choice before you, it seems that your choice and DEATH's choice are importantly and counterfactually linked. Because DEATH knows your mind so well, if you were to choose Aleppo (without the benefit of a truly random coin), he would have chosen it too. Buying the coin allows you to make yourself opaque to DEATH even though your choice comes after DEATH's choice. The upshot here is that CDT is blind to some features of the world that depend on your action.

4 Functional Decision Theory

EDT gets cases like RANDOM COIN right. After all, flipping the random coin is correlated with escaping DEATH and gaining more utility than not flipping the coin. However, this is only because flipping and survival are correlated *in some way*. As we've seen, EDT does not have the apparatus to distinguish between spurious and non-spurious correlations and thereby gets a number of other cases wrong, including XOR BLACKMAIL.

The lesson from RANDOM COIN is that CDT is too parsimonious in what it counts as a non-spurious correlation. In this case, the correlation between DEATH's choice and yours if you do not buy the coin is intuitively robust. If you had not bought the coin and had gone to Aleppo, DEATH certainly would have met you there, at least on a very natural counterfactual.

Because DEATH knows how (or at least bases his decisions on how) you make your decisions, his choice and your choice—barring randomness—are non-spuriously correlated. This is the primary motivation behind functional decision theory. Functional decision theory agrees with CDT that not every correlation between action and state should be decision-relevant. Instead, what matters is counterfactual correlation based on what *depends* on your actions. However, according to FDT, CDT has the wrong view of what should count as a legitimate dependency hypothesis.

In brief, FDT suggests that you should think of yourself as instantiating some decision algorithm. Given certain inputs, that algorithm will output some action. Because algorithms are multiply realizable, however, there can be other instantiations of this same (or a very similar) algorithm, and, more generally, the state of the world can non-causally depend on what your decision algorithm does. Your decision

9. Although this case is a counterexample to the Joyce and Arntzenius versions of CDT, we do not rule out the possibility that an alternative version would recommend buying the random coin in this case. Our case below—DEATH ON OLYMPUS—is a more general counterexample.

procedure is not local to your own mind, and rationality requires taking this fact into account.[10]

One thing we know about mathematical functions is that different instances of the same function do not behave differently on the same input. For example, if you are playing a Prisoner's Dilemma against your psychological twin, you know that you both run the same decision procedure. If your decision procedure were to output *defect*, then your twin's would as well, given that she's computing the same procedure.[11]

One way to spell out the details of Random Coin is in similar terms: Simply stipulate that death runs (as a subroutine) a copy or near copy of whatever procedure you use to decide whether to stay, flee, or randomize. In this way, you can know that whatever your procedure outputs, death's will output the same.

Of course, given the explicit set up of the case, death is not necessarily running a "copy" of you in his head in order to make his prediction. However, regardless of the means he employs, if he's reliable, then his decision will depend on the output of your decision algorithm. For instance, death may simply have an appointment book that either lists your location on any given day or says you will randomize. In this case, death doesn't personally know the ins and outs of your algorithm, but simply looks up what you'll do. According to the fdt counterfactuals developed below, however, if you think the appointment book is a reliable predictor of where you'll be, then what the appointment book itself says depends, in your view, on how your decision procedure behaves. The reason it says "Damascus" is that your decision procedure outputs *stay*. So, since you are free to choose between staying, fleeing, and randomizing, where death ends up depends on your *choice* even though his location is causally independent of your decision as a matter of physical fact.

fdt agents consider only what depends on their decision, but they imagine intervening not on the action directly (as cdt does) but on the algorithm that determines their action. That is, they imagine what would happen differently if they were to come to different verdicts at the end of deliberation.[12]

Let's consider a second example to get a better sense of how this works. Consider figure 3. In Death in Damascus, the output of the agent's decision algorithm determines both her action (S or F), and death's location. If she sets her algorithm to output F, then that changes death's location to Aleppo and results in her fleeing (in the counterfactual world that the agent imagines). Thus, according to that counterfactual, if she fled to Aleppo, she would both die and lose \$1. On the other hand, if she sets her algorithm to output S, then she stays and death ends up in Damascus. According to that counterfactual, if she stayed in Damascus she would die but not lose any money. Since outputting S results in a better outcome, she stays.

Somehow or other the predictions of reliably good predictors depend on the output of your decision function. You are still free to choose between all the actions in the act-space, but because of this mathematical relationship between you and the predictor, her prediction depends on what you decide to do even in the absence of any causal connection between you. Indeed, the prediction could have already taken place before you made your decision or before you were even born. Subjunctive dependence is, in this case, *not time sensitive.*

4.1 Subjunctive Dependency and Logical Counterfactuals

Formally, then, fdt requires agents to have a probability and utility function along with a set of dependency hypotheses. This time, the hypotheses are of the form

$$\bigwedge_{a \in \mathcal{A}} \texttt{MyAlgorithm}(\textbf{input}) = a \mathrel{\square\!\!\rightarrow} s$$

10. We borrow this phrasing from Andrew Critch.

11. We assume the procedure is deterministic here for simplicity.

12. To be clear, like cdt and edt, fdt tells the agent which *actions* to perform, not which algorithm to have. It arrives at the recommendation, however, by considering potential interventions on the agent's algorithm.

where `MyAlgorithm` is the agent's decision algorithm, s is some state of the world, and **input** is a vector of inputs. The **input** vector contains parameters that may affect the agent's decision, such as her probability function, utility function, observations, and so on.

In other words, the agent's dependency hypotheses are conjunctions of counterfactuals about what *would* happen if her decision algorithm on a given input were to output a particular action. Because other agents may have the same algorithm, in the hypothetical scenario where her algorithm outputs a on input $\mathbf{x}$, theirs does too. More broadly, the state of the world can non-causally depend on the output of the agent's algorithm according to FDT's dependency hypotheses.

One worry is that this notion of dependence seems to involve counter-logicals, i.e., questions about what would be the case if a particular function were to produce a different output.[13] This worry is correct. On the best specification of FDT that we know of, counter-logicals or counterpossibles as discussed in Bennett (1974), Cohen (1990), and Bjerring (2013) are required. Indeed, we also need a notion of functional similarity between distinct but related algorithms. If the predictor is running, say, a decent but imperfect approximation of you, then the output of your algorithm and hers should still counterfactually covary.

Unfortunately for us, there is as yet no full theory of counterlogicals or functional similarity, and for FDT to be successful, a more worked out theory is necessary, unless some other specification is forthcoming. Nonetheless, exactly how these counterlogicals work is a topic for another discussion. On all the problems we consider in this paper, it's clear what the dependency hypotheses should be for FDT agents. The most important consideration is that whatever the right theory of logical counterfactuals, multiple instances of the same function move in subjunctive lock-step: if one copy were to output a on input $\mathbf{t}$, then the others would as well.

4.2 FDT's Notion of Expected Utility

Aside from the difference in dependency hypotheses, FDT is structurally very much like CDT.[14] The primary difference is that, instead of advising agents to consider the causal consequences of what would happen if their limbs were to move differently, it instead advises them to consider the consequences of what would happen if their algorithm were to arrive at a different verdict at the end of rational deliberation.

To home in on such a notion, we first consider:

$$U_{\text{FDT}_1}(a) = \sum_{s \in \mathcal{S}} P(\texttt{MyAlgorithm}(\mathbf{t}) = a \mathbin{\square\!\!\rightarrow} s) u(o[a, s])$$

The counterfactual arrow here $\square\!\!\rightarrow$ is now not meant to indicate physical (CDT-style) counterfactuals, but instead to indicate the counterfactual based on her logical dependency hypotheses discussed in the previous section. The $\mathbf{t}$, as before, is a vector of inputs fed into `MyAlgorithm`.

For the sake of specificity, we'll primarily be interested in agents who themselves follow algorithms that comply with FDT's recommendations. Such algorithms take as input a probability function, a causal graph that encodes the agent's views on subjunctive dependence, the agent's utility function, and the set of the agent's observations. Because the latter two parameters will be specified in each decision problem we discuss below, we'll focus on the first two parameters: the probability function P and the graph G. (Here, G is similar to the causal graphs we have already seen, except it also captures logical influence as in figure 3.)

Because the agent we're interested in follows FDT's recommendations, we use $\underline{\text{FDT}(\text{P},\text{G})}$ as a variable indicating the agent's algorithm when fed a full set of

13. For more on why counterlogicals appear to be necessary for a full specification of FDT, see Yudkowsky and Soares 2017.

14. On the authors' favored version of FDT there are in fact a few more important differences, but these need not concern us here. For details, see Dai (2009).

parameters. We then have:

$$U_{\text{FDT}_2}(a) = \sum_{s \in \mathcal{S}} P(\underline{\text{FDT(P,G)}} = a \mathbin{\square\!\!\rightarrow} s) u(o[a, s])$$

Expected utility for the FDT agent is determined by considering what would happen if FDT-style reasoning were to lead to a given act when fed parameters P, G, and so on.

So that we can graphically depict these relationships, we'll again use a Pearl-style formulation of FDT so that our official equation is:[15]

$$U_{\text{FDT}}(a) = \sum_{s \in \mathcal{S}} P(s \mid \texttt{do}(\underline{\text{FDT(P,G)}} = a)) u(o[a, s]) \tag{3}$$

FDT advises agents to perform the act that maximizes equation (3). Note that although FDT tells agents which *act* to perform, the place where the agent intervenes in her causal graph when she is considering her options is the node representing the deliberative process that will determine her act—not the node representing the act itself. That is, the agent thinks about possible future outcomes by considering different possible conclusions her rational deliberation might reach, not by considering different possible actions her body might carry out (though the latter can obviously change *based on* the conclusion she reaches). In a slogan, FDT says: *Intervene on your algorithm, not on your action.*

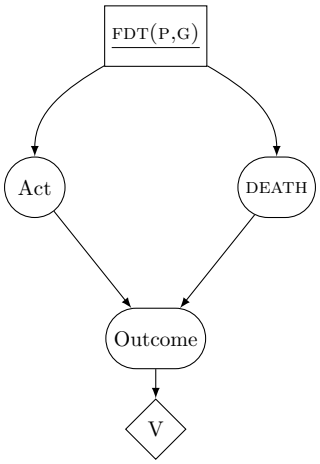


Figure 3: A causal graph for FDT agents facing Death in Damascus, Random Coin, and Death on Olympus. The agent intervenes on $\underline{\text{FDT(P,G)}}$, which controls the Act variable.

4.3 FDT Distinguished from EDT and CDT

FDT, CDT, and EDT all agree in scenarios where every correlation between the agent's action and an event in the world is caused by the agent's action. They disagree

15. As with equation (2) for CDT, equation (3) is not fully general, since the agent may be uncertain over various causal graphs. (See footnote 3.) Therefore, the fully generalized equation for FDT is:

$$U_{\text{FDT}}(a) = \sum_i \sum_{s \in \mathcal{S}} P(G_i) P(s \mid \texttt{do}_i(\underline{\text{FDT(P,G)}} = a)) u(o[a, s])$$

where G_i is a possible graph, and $\texttt{do}_i$ is the $\texttt{do}$-operator for that graph. Again, we avoid this added complexity since the structure of the cases considered in this paper always determines a unique graph.

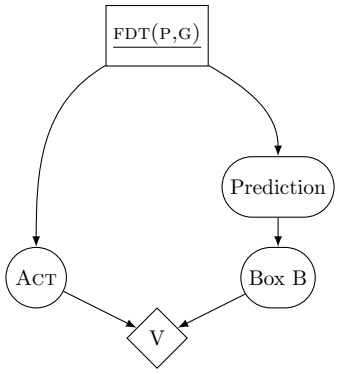


Figure 4: Newcomb for FDT agents. Changing the output of FDT(P,G) changes both the agent's action and the predictor's prediction, which in turn affects whether there is money in the opaque box.

when this assumption is violated, such as in situations where the behavior of an accurate predictor correlates with, but is not caused by, the agent's action. FDT gets different verdicts from CDT in the cases we'll discuss below, but also in more standard cases like Newcomb's problem (Nozick 1969):

Newcomb An agent finds herself standing in front of a transparent box labeled "A" that contains $1,000, and an opaque box labeled "B" that contains either $1,000,000 or $0. A reliable predictor, who has made similar predictions in the past and been correct 99% of the time, claims to have placed $1,000,000 in box B iff she predicted that the agent would only take box B. The predictor has already made her prediction and left. Box B is now already full or already empty. Should the agent take both boxes ("two-box"), or leave the transparent box containing $1,000 behind ("one-box")?

On FDT's view, the predictor decides whether to place money in Box B based on an accurate prediction of the agent. We assume first that the predictor's prediction is dependent in some way upon the agent's algorithm. So, although the predictor's and agent's actions are *causally* independent, they are not *subjunctively* independent according to FDT, since they both are determined by the output of the same (or similar) algorithms. Therefore, $P(\text{full} \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{two-box})) \ll P(\text{full} \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{one-box}))$. So, FDT recommends one-boxing. For the causal graph, see figure 4.

Note, however, that in Newcomb, as elsewhere, *dependency* between the prediction and the output of the algorithm is required. Suppose, for instance, that 99% of blue-eyed people take two-boxes, and the FDT agent is herself blue-eyed. Suppose further that she knows the predictor actually makes her prediction based only on the eye-color of the agent. In that case, the agent will not think there is any dependence between her algorithm's output and the prediction and will instead two-box.[16]

Because FDT gets the same verdict as EDT in the primary cases we'll be focused on, it's important to distinguish it from EDT as well. To that end, we revisit XOR Blackmail from the perspective of an FDT agent. In this case, the blackmailer decides whether to send the letter based on (i) what the agent would do upon receipt, and (ii) whether the agent in fact has termites. Furthermore, the agent knows that the output of her algorithm has an effect on what the predictor does but not on whether her house is infested. If she gets the letter, she reasons: *If I were to pay, then that might change whether I would have gotten this letter, but it would not change whether I have termites. If there are termites and I were to pay, I'd be out a million and a thousand dollars. If there are not, I'd be out a thousand dollars.*

16. FDT theorists, of course, need a more general theory of when a prediction is dependent upon the agent's algorithm as mentioned above. Note, however, that since FDT uses Pearl-style causal graphs, such dependencies can partially be read off from the agent's conditional probabilities.

On the other hand, if I were to avoid paying and there are termites, I would lose a million dollars. If there are not, I'd lose nothing. Therefore, not paying dominates.

More explicitly, in XOR BLACKMAIL, there are four states of the world: (i) the letter is sent, and there are termites (st), (ii) the letter is sent, and there are no termites ($s\bar{t}$), (iii) the letter is not sent, and there are termites ($\bar{s}t$), and (iv) the letter is not sent, and there are no termites ($\overline{st}$).

Payoffs are determined by whether there are termites, and whether the agent pays. Furthermore, although the exact counterfactual probabilities are not determined by the set up, we know that paying does not counterfactually influence whether there are termites. I.e., whether there's in fact a termite infestation does not depend on the output of the agent's algorithm. So, $P(t \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{pay})) = P(t)$ and $P(t \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{don't})) = P(t)$. We then have:

$$\begin{aligned}
U_{\text{FDT}}(\text{pay}) &= P(st \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{pay}))u(st, \text{pay}) \\
&\quad + P(\bar{s}t \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{pay}))u(\bar{s}t, \text{pay}) \\
&\quad + P(s\bar{t} \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{pay}))u(s\bar{t}, \text{pay}) \\
&\quad + P(\overline{st} \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{pay}))u(\overline{st}, \text{pay}) \\
&= P(t \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{pay}))u(t, \text{pay}) \\
&\quad + P(\bar{t} \mid \texttt{do}(\underline{\text{FDT(P,G)}} = \text{pay}))u(\bar{t}, \text{pay}) \\
&= -1,001,000P(t) - 1,000P(\bar{t}) \qquad (4)
\end{aligned}$$

On the other hand, by similar reasoning, we have:

$$\begin{aligned}
U_{\text{FDT}}(\text{don't}) &= P(t \mid (\underline{\text{FDT(P,G)}} = \text{don't})u(t, \text{don't}) \\
&\quad + P(\bar{t} \mid (\underline{\text{FDT(P,G)}} = \text{don't})u(\bar{t}, \text{don't}) \\
&= -1,000,000P(t) - 0P(\bar{t}) \qquad (5)
\end{aligned}$$

Since (4) is always less than (5), not paying is always superior. The effect of paying the blackmail is simply changing the conditions under which the letter is sent, which is not a real improvement in the agent's situation. FDT, then, is a genuinely new theory, since its verdicts diverge both from those of CDT and from those of EDT.

5 Cases

We now consider a variety of unstable cases, in each one contrasting the response of CDT with FDT. Each is designed to illustrate both the virtues of FDT and CDT's failure to recognize the counterfactual link between multiple instances of an algorithm.

5.1 Olympus

First, consider the following case of causal dominance.

Death on Olympus You have three options. You can remain in Damascus, you can travel for free to Aleppo, or you can pay \$1,001 to climb Mount Olympus. DEATH fixes a day when he will try to meet you. He predicts ahead of time where you'll be when you die, and if you are somewhere else then you get to cheat DEATH and live forever. The day before you die, DEATH tells you I AM COMING FOR YOU TOMORROW. You value immortality at \$1,000. However, if you end up climbing Olympus and dying there, you get to speak with the gods post-mortem. Such a conversation is worth \$1,501 to you. So, on net, between the cost of the climb and the chat with the deities, dying on Olympus is worth \$500, but surviving on Olympus is worth −\$1.

		Death		
		Aleppo	Damascus	Olympus
Carl	Aleppo	0	1000	1000
	Damascus	1000	0	1000
	Olympus	−1	−1	500

Table 1: The CDT agent's payoff matrix in DEATH ON OLYMPUS. Because DEATH's location and Carl's choice are causally independent, he considers climbing Mt. Olympus a dominated option.

Suppose DEATH tells Carl, the CDT agent, I AM COMING FOR YOU TOMORROW. What does he do?

Answer: he either remains in Damascus or flees to Aleppo. To see why, first consider Carl's payoff matrix shown in Table 1.

DEATH's location and Carl's location are causally independent. So, from Carl's standpoint, DEATH is either in Aleppo, Damascus, or Olympus, and there's nothing he can do about that. Carl reasons that no matter where DEATH is, both staying in Damascus and fleeing to Aleppo strictly dominate climbing Olympus (according to CDT counterfactuals).

Since choosing dominated options is irrational, Carl will therefore end up in either Aleppo or Damascus, where he is sure to die and receive a payoff of 0. Note that this answer is mandatory for defenders of any standard version of CDT.[17] Although some readers may disagree with Joyce's or Arntzenius's solutions to unstable problems, every version of CDT is committed to denying that going to Olympus is rational because it's dominated by both alternatives. CDT's dependency hypotheses are therefore hopelessly impoverished.

Fiona, the FDT agent, fares much better. For her, DEATH's location counterfactually varies with her own choice. Since DEATH uses the same algorithm she does to predict where she'll be, she thinks that if she were to go to either Aleppo or Damascus, DEATH would be there, and she'd consequently bring about a payoff of 0.[18] On the other hand, if she were to go to Olympus, DEATH would be there, and she'd get a payoff of 500. So Fiona packs her gear, ponders the questions she'll ask the gods, and begins her climb.

Both EDT and FDT arrive at the right answer, but only FDT does so for the right reasons. When Fiona sets out for Olympus, that is good evidence that DEATH will be there, which is why EDT recommends climbing. What FDT sees is that this is a special kind of evidence whose link to DEATH's location is counterfactually robust. Where Fiona decides to go determines where DEATH goes, but for logical and not for causal reasons. DEATH effectively has a copy of Fiona that he runs in his mind before making his choice, and multiple copies of the same program are sure to get the same result. CDT misses this dependency entirely, and EDT considers it just another kind of correlation.

5.2 Psychopath Button

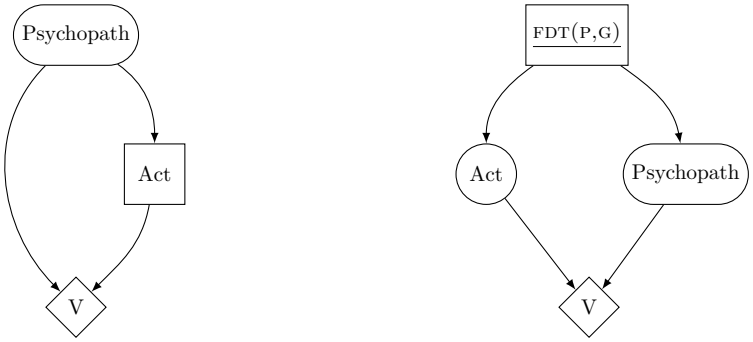
We now turn to a much-discussed example from Egan (2007):

Psychopath Button You are debating whether to press the "kill all psychopaths" button. It would, you think, be much better to live in a world with no

17. The one possible exception we're aware of is Wolfgang Spohn. 2012. "Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box." *Synthese* 187 (1): 95–122, which defends a very non-standard version of CDT (that, e.g., prescribes taking one box on Newcomb's problem).

18. As before, DEATH may make his decision in some way other than by running a copy of her algorithm but that nonetheless depends on her algorithm. Since these alternatives won't affect FDT's verdicts in the cases here and below, we will ignore them henceforth.

psychopaths. Unfortunately, you are quite confident that only a psychopath would press such a button. You very strongly prefer living in a world with psychopaths to dying.



(a) A graph for CDT agents.

(b) A graph for FDT agents

Figure 5: A graph of PSYCHOPATH BUTTON for CDT and FDT agents.

From the CDT point of view (represented in Figure 5a), whether you're a psychopath or not is simply a feature of the environment you cannot change. However, it influences both your actions (whether you press the button), and the outcome.

Although we did not plug in exact utilities this time, under a Joycean solution, CDT seems to get this case close to right (*pace* Egan). The more confident Carl is that he'll press the button, the more confident he becomes he's a psychopath. In turn, he becomes more attracted to not pressing the button because he strongly wishes not to die. This strong preference leads him to an equilibrium probability with high confidence that he will not, in fact, press the button. (Though he still may toss a heavily biased coin and push the button with small probability.)

It's less clear how we should model this case from the point of view of FDT, and there are a variety of options.[19] The most natural and illustrative, we think, is to assume that what actions you would or would not perform in various (hypothetical or real) circumstances determines whether you're a psychopath. What actions you would perform in which circumstances is in turn determined by your decision algorithm. On this reading of this case, the potential outputs of your decision algorithm affect both whether you'd press the button *and* whether you're a psychopath.

In practice, this leads the FDT agent *always* to refrain from pressing the button, but for very different reasons from the CDT agent. Fiona reasons that if she *were* to press the button that kills so many people, then she would be a psychopath. She does not regard her psychopathic tendencies—or lack thereof—as a fixed state of the world isolated from what decision she actually makes here.

This seems like the right reasoning, at least on this understanding of what psychopathy is. Psychopaths just are people who tend to act (or would act) in certain kinds of ways in certain kinds of circumstances. We can take for granted that everyone is either born a psychopath or born a non-psychopath, and that Fiona's action cannot causally change this condition she was born with. Yet if this condition consists in dispositions to behave in certain ways, then whether Fiona is a psychopath is subjunctively tied to the decisions she actually makes. If you would not perform A in circumstances C, then you also would not be the *kind* of person who performs actions like A in circumstances like C. FDT vindicates just this sort of reasoning, and refrains from pressing the button for the intuitively simple reason of "if I pressed it, I'd be a psychopath" (without any need for complex and laborious

19. Our interpretation of MURDER LESION below is effectively another way of modeling PSYCHOPATH BUTTON for FDT. On that alternative, whether you are a psychopath is not determined by the cognitive algorithm you use to make decisions, but instead routes around your reasoning process to (partially) control your actions directly.

ratification procedures). When we intervene on the value of the FDT(P,G) variable, we change not just what you actually do, but also what kind of person you are.

5.3 Murder Lesion

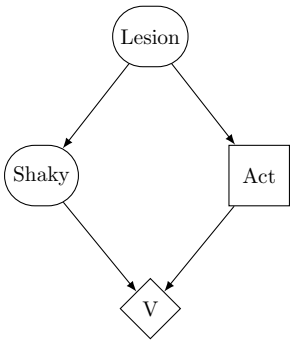


Figure 6: A graph of MURDER LESION for CDT agents.

Finally, consider the following case discussed at length in Joyce (2012):

Murder Lesion Life in your country would be better if you killed the despot Alfred. You have a gun aimed at his head and are deciding whether to shoot. You have no moral qualms about killing; your sole concern is whether shooting Alfred will leave your fellow citizens better off. Of course, not everyone has the nerve to pull the trigger, and even those who do sometimes miss. By shooting and missing you would anger Alfred and cause him to make life in your country much worse. But, if you shoot and aim true the Crown Prince will ascend to the throne and life in your country will improve. Your situation is complicated by the fact that you are a random member of a population in which 20% of people have a brain lesion that both fortifies their nerve and causes their hands to tremble when they shoot. Eight in ten people who have the lesion can bring themselves to shoot, but they invariably miss. Those who lack the lesion shoot only one time in 10, but always hit their targets.

For definiteness, assume that the payoffs are as described in Table 2, with S denoting the act of shooting, and L denoting the state of having the lesion.

	L	$\neg L$
S	-30	10
$\neg S$	0	0

Table 2: Payoff matrix for MURDER LESION, where S refers to the action of shooting, and L is the proposition that you have the lesion.

The story for CDT agents is familiar. The causal graph for CDT agents is depicted in Figure 6, and this case is equivalent to DEATH IN DAMASCUS for such agents with only variables and numerical values changed. If you're confident you'll shoot, you end up evaluating $\neg S$ more favorably, which in turn provides reason not to shoot. You then lower your credence you'll shoot, eventually achieving equilibrium when $U_e(S) = U_e(\neg S) = 0$. In this particular case, $P(L \mid \mathtt{do}(S)) = P(L \mid \mathtt{do}(\neg S)) = P(L)$ because the Act-node is a descendant of the Lesion-node. So, the expected utility of shooting and not shooting are respectively calculated as follows:

$$\begin{aligned} U(S) &= P(L)u(S, L) + P(\neg L)u(S, \neg L) \\ U(\neg S) &= P(L)u(\neg S, L) + P(\neg L)u(\neg S, \neg L) \end{aligned}$$

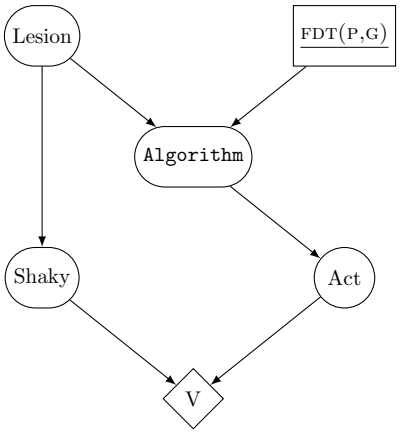


Figure 7: The graph for FDT agents facing MURDER LESION. Both the lesion and the FDT algorithm partially determine which algorithm the agent actually uses to decide.

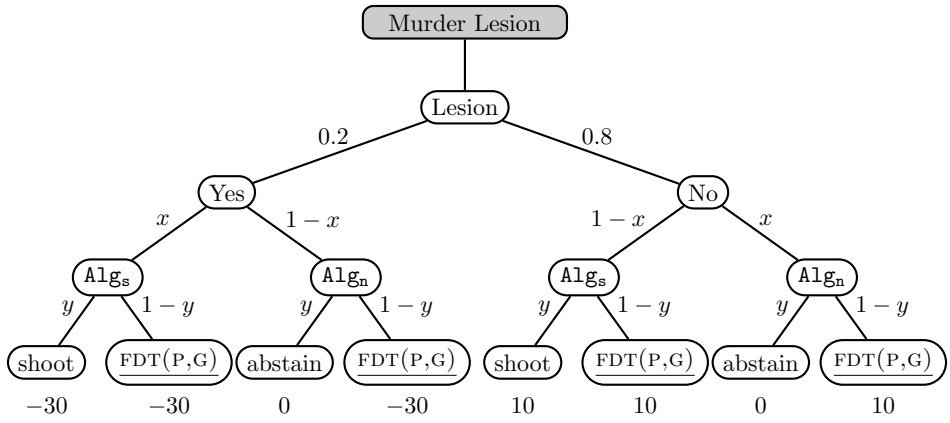


Figure 8: One version of MURDER LESION, where FDT will advise agents to shoot. Whether the agent has the lesion determines how likely she is to be pre-wired with either $\texttt{Alg}_\texttt{s}$ or $\texttt{Alg}_\texttt{n}$, which in turn determines how likely she is to either shoot automatically, abstain automatically, or follow the verdict of FDT. The conditional probability of moving from a node to a child is given on the edge between them. The payoffs of each leaf are listed below it.

In equilibrium (i.e., for P_e), the two equations both take on the same value of 0 and $P_e(L) = .25$.

However, note that once we've achieved equilibrium, the correlation between S and L is no longer decision-relevant. That is, even though $P_e(L \mid S)$ is not necessarily equal to $P_e(L)$, this stochastic relationship has no bearing on the agent's decision.

What this means is that *no matter how strong the correlation between shooting and shaky hands,* CDT agents will shoot with 25% probability. To bring out the severity of the problem, imagine that the correlation were near perfect. Conditional on shooting, you're almost certain to have the lesion and miss, and conditional on not shooting, you're almost certain not to have the lesion but have steady hands. In this case, not shooting guarantees a payoff of 0, while shooting guarantees a payoff of near −30. Nonetheless, CDT agents sometimes still shoot and on average receive a payoff of −7.5. That CDT agents do predictably worse than agents who never shoot in cases like this seems irrational.

It is not immediately clear how FDT handles this problem. One might think that FDT's answer could be calculated by a graph like Figure 6, with Act replaced by $\underline{\text{FDT(P,G)}}$. If so, then it would reach the same answer as CDT, and the fact that FDT mishandles this problem would call into question the diagnosis that CDT's fundamental problem is ignoring subjunctive dependence.

However, using that graph would set up a contradiction with the problem statement. As defined here, $\underline{\text{FDT(P,G)}}$ is a deterministic function. Once it is fed a causal graph, utility function, and probability distribution, its output is fully determined. But the arrow between the lesion and $\underline{\text{FDT(P,G)}}$ would imply that the presence or absence of the lesion affects the output of the (fixed) FDT equations, which is false. In other words, the Lesion-node cannot be an ancestor of the $\underline{\text{FDT(P,G)}}$-node. Therefore, in order for it to be possible for the graph to comply with the case description, we require that the agent's action not be fully determined by the output of the $\underline{\text{FDT(P,G)}}$ algorithm alone.

At this point, the official description of the case leaves underspecified exactly how the lesion and FDT jointly end up affecting the action. Thus *the case is underspecified from the point of view of the functional decision theorist,* which explains the initial lack of clarity. Critically, this is not true of a causal decision theorist who accepts Figure 6 as a full specification, which she might do because she believes she can intervene directly on actions.

One Specification of Murder Lesion

We will spell out one of many possible ways of filling in the details, which we hope will be illuminating. However, different formulations of the details can yield different problems, and thus different correct solutions.

Assume that the lesion affects two things: whether the agent's hands are shaky, and what the agent's actual decision algorithm is. To make the case interesting, we'll assume that FDT gets some say as well. Its recommendation at least sometimes determines what the agent's actual decision algorithm does. Figure 7 provides a causal graph of this scenario.

Suppose that as before, 20% of the population have the lesion and 80% do not. If you have the lesion, you come hard-wired with one of two algorithms: $\texttt{Alg}_\texttt{s}$ or $\texttt{Alg}_\texttt{n}$. $\texttt{Alg}_\texttt{s}$ automatically returns `shoot` with probability y regardless of what FDT says to do. Similarly, $\texttt{Alg}_\texttt{n}$ automatically returns `abstain` with probability y regardless of FDT's advice. With probability $1 - y$, both algorithms return whatever FDT says to do. Given the presence of the lesion, there's an x chance $\texttt{Alg}_\texttt{s}$ directly determines your action, and a $1 - x$ chance $\texttt{Alg}_\texttt{n}$ does. Given the absence of the lesion, there's an x chance $\texttt{Alg}_\texttt{n}$ controls your actions, and a $1 - x$ chance $\texttt{Alg}_\texttt{s}$ does. (See Figure 8.)

Under this specification, the FDT algorithm is uncertain about whether the agent has the lesion or not. Since y proportion of the time, the agent will not listen to FDT regardless of what it says (because she is an auto-shooter or auto-abstainer), those cases do not factor into FDT's deliberations. The remaining $1 - y$ proportion

of the time, FDT *does* have control over the agent's actions, in which case outputting `abstain` adds 0 utility, but outputting `shoot` adds $2 - 2y$ utility in expectation. FDT therefore outputs `shoot` in this scenario. The higher y, the less likely FDT has control, and the less FDT can contribute in expectation to her payoff.

This seems to us like the correct behavior. Insofar as the lesion prevents FDT from affecting the agent's actions, the decision of FDT is irrelevant. On the other hand, insofar as FDT does control the agent's actions, there's no reason to take into account any correlation between the presence or absence of the lesion and FDT's output, since that connection is rightly severed when we condition on the action being determined by FDT. So, insofar as FDT has control over what you do, it should only take into account the initial probability that you have the lesion, the graphical structure, and how much better or worse it would be to kill, miss, or abstain from shooting.

There are, of course, other ways of interpreting MURDER LESION. However, as far as we can tell, none create a problem for FDT. Although *agents* who shoot may end up worse off on average than agents who do not, this is only because of factors outside of FDT's control and invariant under the actions a decision theory should recommend.

6 Conclusion

CDT's ability to distinguish causal links from mere correlations in its decision-making is rightly recognized as an improvement over EDT. Not every correlation between act and state is decision-relevant. Instead, rational agents should only take into account potential ways in which the world depends on the act they perform.

Unfortunately, CDT is blind to non-causal dependencies that can be crucial to rational decision-making. When predictors have access to your decision procedure, their prediction is logically at least partially determined by your action. That is because multiple instantiations of the same function will return the same output on the same input.

We sketch a theory of rational action, FDT, to be fully developed over the course of a series of papers, which uses this idea of subjunctive dependency to correctly prescribe rational actions across a wide range of cases.

The cases we used to argue for FDT and against CDT had, to our minds, relatively unambiguously correct verdicts. These provide strong data in favor of the former theory and, in turn, for the decision-relevance of logical dependencies. FDT also diverges from CDT in more contentious cases like NEWCOMB. Given that intuitions are strongly divided on that case, it cannot be counted as dispositive. However, if FDT becomes the consensus theory, one-boxing should also become the consensus strategy.

As we noted, FDT still faces a number of foundational issues. Regardless of how FDT stacks up against CDT and EDT, it remains the case that we need some account of reasoning under uncertainty about tautologies and contradictions or some reformulation of FDT that avoids these problems. Without such an account, FDT will remain underspecified in important respects. These questions appear ripe for further investigation.

References

Ahmed, Arif. 2014. "Dicing With Death." *Analysis* 74 (4): 587–592.

Arntzenius, Frank. 2008. "No Regrets, or: Edith Piaf Revamps Decision Theory." *Erkenntnis* 68 (2): 277–297.

Bennett, Jonathan. 1974. "Counterfactuals And Possible Worlds." *Canadian Journal of Philosophy* 4 (2): 381–402.

Bjerring, Jens Christian. 2013. "On Counterpossibles." *Philosophical Studies,* no. 2: 1–27.

Cohen, Daniel. 1990. "On What Cannot Be." In *Truth or Consequences: Essays in Honor of Nuel Belnap,* 123–132. Kluwer Academic Publishers.

Dai, Wei. 2009. "Towards a New Decision Theory." *Less Wrong* (blog), August 13. `http://lesswrong.com/lw/15m/towards_a_new_decision_theory/`.

Eells, Ellery. 1984. "Newcomb's Many Solutions." *Theory and Decision* 16 (1): 59–105.

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116 (1): 93–114.

Gauthier, David. 1994. "Assure and Threaten." *Ethics* 104 (4): 690–721.

Gibbard, Allan, and William L. Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility: Theoretical Foundations." In *Foundations and Applications of Decision Theory,* edited by Clifford Alan Hooker, James J. Leach, and Edward F. McClennen, vol. 1. The Western Ontario Series in Philosophy of Science 13. Boston: D. Reidel.

Jeffrey, Richard C. 1983. *The Logic of Decision.* 2nd ed. Chicago: Chicago University Press.

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

———. 2012. "Regret and Instability in Causal Decision Theory." *Synthese* 187 (1): 123–145.

Lewis, David. 1981. "Causal Decision Theory." *Australasian Journal of Philosophy* 59 (1): 5–30.

Meacham, Christopher J. G. 2010. "Binding and its Consequences." *Philosophical studies* 149 (1): 49–71.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday,* edited by Nicholas Rescher, 114–146. Synthese Library 24. Dordrecht, The Netherlands: D. Reidel.

Pearl, Judea. 1996. "Causation, Action, and Counterfactuals." In *6th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-'96),* edited by Yoav Shoham, 51–73. San Francisco, CA: Morgan Kaufmann.

———. 2009. *Causality: Models, Reasoning, and Inference.* 2nd ed. New York: Cambridge University Press.

Savage, Leonard J. 1972. *The Foundations of Statistics.* Dover books on Mathematics. Dover Publications.

Skyrms, Brian. 1980. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws.* New Haven, CT: Yale University Press.

———. 1990. *The Dynamics of Rational Deliberation.* Cambridge, UK: Cambridge University Press.

Soares, Nate, and Benja Fallenstein. 2015. "Toward Idealized Decision Theory." arXiv: `1507.01986 [cs.AI]`.

Spohn, Wolfgang. 2012. "Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box." *Synthese* 187 (1): 95–122.

Yudkowsky, Eliezer. 2010. *Timeless Decision Theory.* The Singularity Institute, San Francisco, CA. `http://intelligence.org/files/TDT.pdf`.

Yudkowsky, Eliezer, and Nate Soares. 2017. "Functional Decision Theory: A New Theory of Instrumental Rationality." arXiv: `1710.05060 [cs.AI]`.