Tournament Decision Theory

The dispute in philosophical decision theory between causalists and evidentialists remains unsettled. Many are attracted to the causal view's endorsement of a species of dominance reasoning, and to the intuitive verdicts it gets on a range of cases with the structure of the infamous Newcomb's Problem. But it also faces a rising wave of purported counterexamples and theoretical challenges. In this paper I will describe a novel decision theory which saves what is appealing about the causal view while avoiding its most worrying objections, and which promises to generalize to solve a set of related problems in other normative domains.

Our path towards this view begins with a diagnosis of the problems facing causal decision theory as a special case of a more general phenomenon called *decision-dependence* – the possibility that some crucial input into our evaluation of options is affected by the decision we end up making. A wide range of attractive views in ethics and practical rationality run into cases of decision-dependence, and they will all face problems analogous to those troubling the causalist. We will go on to look at a promising initial thought about how to approach simple two-option decisions in these cases that I believe is on the right track, and see why attempts to generalize it to decisions with three or more options seem doomed to fall apart.

The solution will be a radically revised approach to decision-making, *Tournament Decision Theory*, which models decision problems after the structure of a tournament, in which each option competes with each other option pairwise, and the winner is determined by the overall results. I will show how Tournament Decision Theory both provides us with tools to address the distinctive puzzles raised by decision-dependence, and allows us to claim new ground in the familiar battle between causalists and evidentialists.

I. The Virtues and Vices of Causal Decision Theory

Causal decision theory (henceforth, CDT) is a view about what we ought to do in the subjective sense – that is, in light of our uncertainty about the world. In its most basic form, the idea is the following: consider the possible *causal situations* you might be in, where a causal situation is a state of the world outside your control that determines the effects of your actions. Then for each causal situation, figure out a) how *probable* that situation is by your current lights, and b) the *utility* of the outcome each action would cause in that situation. What you ought to do is select the action which has the highest *causal expected utility*, which is an average of the utility of the outcome it brings about in each causal situation weighted by the probability of that causal situation. More formally, CDT says:

CDT: An action is permissible iff it maximizes causal expected utility (CEU), given by:

$$\mathrm{CEU(X)} = \sum_{i=1}^{n} \mathrm{Pr_0(C_i)U(C_i \,\&\, X)}$$

Where $C_i \ldots C_n$ is a partition of causal situations, $Pr_0(C_i)$ is the agent's initial credence in C_i and $U(C_i \,\&\, X)$ is the utility of the outcome resulting from X being performed in C_i.[1]

The standard foil for CDT is Evidential Decision Theory (henceforth, EDT). The idea behind EDT is that you ought to consider, for each action, how probable it is that each possible outcome will obtain, *conditional* on your performing that action. This amounts to hypothetically adding to your current evidence the proposition that you perform that action, and updating the rest of your beliefs accordingly. The average value of the possible outcomes of the action weighed by this conditional probability gives you the action's *evidential expected utility*, and EDT instructs you to choose the option that ranks highest. More formally, EDT says:

EDT: An action is permissible if it maximizes evidential expected utility (EEU), given by:

$$\mathrm{EEU(X)} = \sum_{i=1}^{n} \mathrm{Pr(C_i \mid X)U(C_i \,\&\, X)}$$

[1] Not all authors agree about the precise formulation of causal decision theory. Perhaps most notably, Joyce (2012) describes his more elaborate view, which does not give the *initial* credence the same role, as a form of CDT. We will discuss Joyce's view later, but if the reader is worried, we can understand this view as "simple" or "standard" CDT, to distinguish it from other views which sometimes share the label.

Where $C_i \ldots C_n$ is a partition of causal situations, $Pr(C_i \mid X)$ is the agent's credence in C_i conditional on X, and $U(C_i \& X)$ is the utility of the outcome resulting from X being performed in C_i.

The technical difference between the two is that CDT uses your initial *unconditional* credences, while EDT uses your *conditional* credences. The deeper philosophical difference, roughly, is that CDT tells you to choose actions insofar as they are likely (by your current lights) to have better *effects*, and EDT tells you to choose actions insofar as they are *evidence* for better outcomes. Of course, in most ordinary cases, your action is evidence for better outcomes precisely *because* it is likely to have better effects. But there are well-known cases where these come apart and the two theories give opposing recommendations.

Newcomb's Problem

The most famous of these is Newcomb's problem. An extremely reliable predictor offers you a choice between taking an opaque box, in which they have earlier placed either one million dollars or nothing at all, or taking both the opaque box and a transparent box containing one thousand dollars. The twist is that the predictor has placed a million dollars in the opaque box if and only if they predicted that would take just the one box.

A payoff matrix (with payoffs in thousands of dollars, corresponding to utility) would look like this:

	Box Full	**Box Empty**
One Box	1000	0
Two Box	1001	1

CDT tells you to take two boxes, because this option *causally dominates* – it has a better outcome in each possible causal situation, and so its CEU is guaranteed to be higher, no matter how likely the causal situations are. EDT tells you to take one box, because while taking one box doesn't *cause* you to end up with more money, it is *evidence* that you will end up with more money.

Lay intuitions about Newcomb's problem are notoriously split. But two-boxing has come to be the favored response of most decision theorists. Many philosophers have been convinced that one-

boxing, and EDT more generally, is committed to an irrational policy of "managing the news" (Lewis 1981) – acting in a way that gives you good news about the world beyond your control. This is bolstered by consideration of structurally similar cases (such as 'medical Newcomb problems' (Price 1991)) where intuitions are unequivocally in favor of CDT. For our purposes, the thing to note is just that CDT's recommendations in Newcomblike cases are a major attraction of the view for its adherents. So if we want to keep what is appealing about CDT, we will want to preserve this advice.

Death in Damascus, Action-Credences, and Instability

In Newcomb's problem, because one option was causally dominant, it did not matter what the agent's initial unconditional credences in the various causal situations were. However, in other cases where conditional and unconditional credences differ and there is *no* dominant option, the initial credences do matter, and this generates serious problems for CDT. Gibbard & Harper's (1976) *Death in Damascus* is a simple example.

Imagine you learn that Death's appointment book has a meeting scheduled with you tomorrow, which you would much prefer to avoid. You know that Death, who is an extremely reliable prognosticator, made a prediction in advance about where you would be, and will show up where he predicted. You can either stay in Aleppo, or go to Damascus. What should you do?

	Death in Damascus	**Death in Aleppo**
Stay in Aleppo	1000	0
Go to Damascus	0	1000

EDT tells you that your choice is a wash. Conditional on your going to Damascus, you're practically certain to run into Death in Damascus, and conditional on your staying in Aleppo, you're practically certain to run into Death in Aleppo. These outcomes are equally bad, so no option is better than the other. This result, while dispiriting, seems intuitively right.

What does CDT recommend? Well, it depends on the initial unconditional probability you assign to Death being in Aleppo and Damascus respectively. If, at the outset, you think it's more likely that Death is in Damascus, CDT recommends you stay in Aleppo, and vice versa. This doesn't match the intuition that the options are tied. But it also raises additional theoretical difficulties. Because Death is a reliable predictor, the probability that Death is in Damascus is epistemically tied up with the

probability that you decide to go to Damascus. So your unconditional credence about Death's location, which you need in order to make a decision, at least implicitly commits you to certain credences about your own actions.

According to a strong version of a thesis sometimes called "Deliberation Crowds Out Prediction" (Rabinowicz 2002), an agent cannot, while engaged in genuine deliberation about what to do, have credences about their performance of the actions presently available to them (their *action-credences*).[2] Different reasons might be given for this – that it destroys the distinction between actions and states that underlies decision theory (Spohn 1977), that it is inconsistent with the agent seeing themselves as genuinely free (Levi 1993) or having agency (Louise 2009), that it is incompatible with the relationship between subjective probabilities and betting rates (Spohn 1977, Levi 1993), or that an agent cannot have sufficient evidence during deliberation to form an action-credence (Price 2007). If any of these are compelling, this threatens the ability of CDT to recommend anything in these cases.[3]

Even some of those who reject the claim that agents shouldn't *have* action-credences during deliberation maintain that these credences should not matter for what an agent ought to do. Joyce (2002) argues that because in a decision-making context, our action-credences are ultimately fully up to us and self-fulfilling, neither our initial action-credences nor the evidence supporting them should be decisive in making it the case that we ought to do one thing rather than another.[4] For example, an agent who is aware that they are lazy and likely to decide not to leave Aleppo should not take this as grounds for going to Damascus. After all, the agent can create their own conclusive evidence to the contrary precisely by making their choice.

CDT's problems do not end there. *Death in Damascus* is an example of a decision that is in an important sense self-defeating. The information that one will do any particular action in this case is enough to guarantee that that action is wrong by CDT's lights. From the perspective of someone who decides to go to Damascus, Aleppo is better, and from the perspective of someone who decides to

[2] On some accounts of what conditional credences are, you cannot have credences conditional on your actions without having unconditional credences in your actions, which would render this thesis problematic for EDT as well. But the unconditional credences don't play any role in choice for EDT, and so it is open to the evidentialist to weaken this connection, for instance by allowing gaps in one's unconditional credences as long as there is some extension of them which would be coherent with their conditional credences. This way out is suggested by Spohn (1978, pg. 77). See also Hajek (2003).

[3] See Hajek (2016) for a case against this thesis.

[4] We should be a bit careful here – doubts about whether one would *follow through* with a decision might be relevant. But here the belief that's doing work is about *the decision itself* – we can postulate that the agent has no doubts about whether she would follow through.

stay in Aleppo, Damascus is better. When an action looks bad by a decision procedure's lights on the supposition that it is chosen, it is *unratifiable*. CDT allows for situations where no action is ratifiable, which raises a second set of worries.

As the agent comes closer and closer to deciding to stay in Aleppo, it seems like their credence that they will end up in Aleppo should rise. So just before deciding to stay in Aleppo, they have reason to think that their action is a causal disaster relative to going to Damascus; this should make them change their mind. But of course as they go in the opposite direction the same problem arises. For sufficiently reflective agents, then, CDT's recommended decision seems *unstable*.

Going through with an action which looks wrong at the point of decision, strikes many as irrational. This suggests that a theory of decision should have a *ratifiability constraint*: One ought not perform an action that looks wrong by the theory's lights conditional on it being chosen. CDT violates this constraint.

CDT, then, has a counterintuitive verdict in *Death in Damascus*, and requires the agent to use credences about their own decision in an illegitimate way in deliberation. We can contrast this to EDT, which gets the case right, and according to which the agent's unconditional action-credences play no decisive role.

Variations on CDT

In response to the problems above, it is natural to look for modest revisions we might make to the simple version of CDT and see if the objections can be avoided. I will describe a few such proposals, before showing that all of them share vulnerability to other counterexamples. The first is the simplest – we could just add ratifiability as a constraint to CDT, leading to:

Ratifiable CDT: An action is permissible if it maximizes causal expected utility and is ratifiable.

Because of cases like *Death in Damascus*, *Ratifiable CDT* entails there are rational dilemmas – when no action is ratifiable, no action will be permissible. Perhaps this is not much more counterintuitive a verdict in that case than the suggestion that both are permissible. But if we want our decision theory to always provide the agent with actionable advice, we might try and avoid dilemmas in the following way:

Permissive Ratifiable CDT: An action is permissible if a) it maximizes causal expected utility and is ratifiable, OR b) no option is ratifiable.

A more sophisticated strategy has been proposed by Arntzenius (2008) and Joyce (2012). According to their view, one should not, as CDT suggests, hold fixed one's initial beliefs about the causal situations, evaluate, and then settle conclusively on the best option according to that evaluation. Rather, one evaluates in the causalist way and then shifts one's action-credences in the direction of the more attractive option, continuing to reevaluate and correct until a kind of action-credence equilibrium is reached, where no further shifting of action-credences is warranted by the evaluation. In ordinary cases, this equilibrium will be reached at full confidence in some action or other, but in cases like *Death in Damascus*, this process will conclude with the agent having their credence split between multiple actions (in particular, with the agent having 50% credence that they will stay in Aleppo and 50% credence they will go to Damascus). According to Arntzenius and Joyce, reaching this sort of equilibrium is all there is to their rational deliberation.

One concern about this strategy is about the relationship between the action-credence equilibrium it recommends and the question we expect a decision theory to answer – namely, what the agent ought to decide or do. It is tempting to think of the view as proposing that the agent pick a third option - perhaps flip a coin in their head and go to Damascus if and only if it comes up tails. But this is not quite right, because such an option may not be available to them (for instance, if they are not capable of randomizing this way), and if it is, it might be something which carries its own decisive costs or penalties.[5] Rather, the view is that the rational part of deliberation ends, not with an intention to randomize, but with a distribution of credence over what one will end up doing.

Isn't this changing the subject, from a practical question to an epistemic one? One can avoid this challenge if one pairs this view, as Joyce does (2002), with the kind of picture defended by David Velleman (1999), according to which a decision just *is* the formation of a certain kind of self-fulfilling belief about what one will do. Then one is able to say that the equilibrium credence is itself a kind of decision, and maintain that the ambitions of a genuinely practical theory have been met.

Of course, this is already taking on more baggage than one might like. Nevertheless, this *Deliberational CDT*, if successful, could solve a number of the problems above. Because the agent's

[5] Thanks to an anonymous reviewer for pressing this point.

action-credences are not subject, on this picture, to initial evidence, and each evaluation is provisional until the equilibrium is reached, the agent could not, in our example, cite preexisting evidence that they will stay in Aleppo as conclusive grounds for going to Damascus. And the introduction of "mixed" decisions consisting of partial credence in multiple actions allows the view to have something to recommend in cases where every pure option is unratifiable. But as we'll see, whether or not they give satisfying verdicts on the basic *Death in Damascus*, neither the deliberational gambit nor the simpler ways to bake in ratifiability get the intuitive results in a number of closely related counterexamples to CDT.

Asymmetric Death in Damascus

One kind of counterexample, pressed forcefully by Andy Egan (2007), takes the form of *Asymmetric Death in Damascus*. This example is just like Death in Damascus except you also have a strong fondness for the city of Aleppo itself, where all your friends and family live. If you were going to die, you'd much rather die in Aleppo, surrounded by people you care about, with the opportunity to say your proper goodbyes, than suffer a tough ride through the desert to Damascus and perish alone. The decision table might look something like this:

	Death in Aleppo	**Death in Damascus**
Stay in Aleppo	100	1100
Go to Damascus	1000	0

Here, the intuitive response is that there is one uniquely permissible option – staying in Aleppo. And this is what EDT recommends – your evidential expected utility for staying is 100, and your evidential expected utility for going to Damascus is 0.

But *none* of the variants on CDT accommodate this. Basic CDT implies that it is permissible to stay in Aleppo if your initial credence that you will stay in Aleppo is 55% or less. Ratifiable CDT implies no action is permissible. Permissive Ratifiable CDT implies every action is permissible.

Deliberational CDT implies that the agent should adopt the mixed credence: staying in Aleppo with probability 55% and going to Damascus with probability 45%.[6]

Dicing with Death

Another variant is due to Ahmed (2014). This case is again just like *Death in Damascus*, except that the agent has a third option: they may, for a small fee, make their decision on the basis of the random flip of a magic coin whose outcome Death cannot predict, staying in Aleppo if it comes up heads, and going to Damascus if it comes up tails.

	Death in Aleppo, Heads	**Death in Aleppo, Tails**	**Death in Damascus, Heads**	**Death in Damascus, Tails**
Stay in Aleppo	0	0	1000	1000
Go to Damascus	1000	1000	0	0
Coinflip	-1	999	999	-1

The intuitive response is that the agent ought to buy the coin. As Ahmed puts it, it is better to play high-stakes hide-and-seek against someone who cannot predict where you are than someone who can. Again, EDT gets this right. The evidential expected utility of choosing either Aleppo or Damascus is 0. The evidential expected utility of buying the coin is 499.

And again, none of the variants of CDT get this right. This is because no matter what the agent's unconditional credence in each causal situation, at least one option will have a higher causal expected utility than buying the coin. If the agent's overall credence that Death is in Aleppo (the sum of Death in Aleppo, Heads and Death in Aleppo, Tails) is higher than their credence that Death is in Damascus, then going to Damascus has a higher causal expected utility. If their credence that Death is in Damascus is higher, then going to Aleppo has a higher causal expected utility. And if their credence is evenly split, then both options have a higher causal expected utility. CDT, therefore, implies they should not buy the coin. Ratifiable CDT implies every option is impermissible. Permissive

[6] Egan (2007) describes several apparent counterexamples to causal decision theory with this structure, including the Psychopath Button and Murder Lesion. If the reader is not moved by the case here, these other examples are even more forceful, by making the asymmetry even more stark.

Ratifiable CDT implies every option is permissible. And Deliberational CDT implies the agent should stick with adopting the mixed credence: staying in Aleppo with probability 50% and going to Damascus with probability 50%.

Taking Stock

Let us pause to summarize the dialectical situation. CDT is attractive to its adherents because it advises two-boxing in Newcomb's problem. But it comes with serious intuitive and theoretical costs which modest revisions at best only partially address. These costs might be sufficient to drive someone towards EDT, which avoids these additional costs – judgments on Newcomb's problem have always been a bit uncertain, after all. What we would really like is a view that could preserve CDT's virtues without its costs – a view that recommends two-boxing, but *doesn't* require us to use our initial action-credences in deliberation, avoids dilemmas, respects intuitively relevant asymmetries, and has plausible verdicts in the cases discussed above. The rest of this paper is dedicated to developing such a view.

II. Decision-Dependence

Before we continue, I want to step back for a moment and consider normative theories of action more generally. Whether they are theories of rationality, prudence, or morality, and whether they are theories of what we ought to do in the objective sense (given all the facts) or the subjective sense (given our beliefs or evidence), these theories generally provide some way of *evaluating* our options – scoring or comparing them along some dimension of attractiveness.[7] Utilitarians about the objective ought, for example, evaluate options according to the happiness that they would produce. Causal and evidential decision theory also come with a way to evaluate our options – by their causal and evidential expected utility, respectively.

Some of these evaluations have a feature I'll call *decision-dependence* – something crucial to determining the evaluation can depend on, or be affected by, the decision itself. The objective utilitarian evaluation is not decision-dependent. The facts about how much happiness any particular option would produce are the same, whether one ends up choosing that option or not. But take a view about the objective ought which evaluates options in part by how good they are for your children, and imagine you face a decision about which of two children to adopt. If you adopt child A, the evaluation

[7] This evaluation may be cardinal or ordinal, depending on whether it carries information about *how much* more attractive one option is than another. In this paper, I will focus on examples of cardinal evaluations.

rests on how good things are for child A, and if you adopt child B, the evaluation rests on how good things are for child B. This kind of evaluation is decision-dependent: an important input into the evaluation (the identity of your child) depends on which action you perform. Other examples of decision-dependent views are versions of so-called "person-affecting" views that evaluate outcomes based on how good they are for the people who exist, or views that evaluate outcomes based on how well they satisfy the desires or projects the agent will have over the course of their life. The identity of the people who exist and the nature of one's future desires or projects can both depend on what action one performs now.[8]

The aforementioned dependences are *metaphysical* – your decision is what *makes it the case* that you have one child rather than another. But evaluations, particularly those relevant to the subjective ought, might also be decision-dependent *epistemically* – some doxastic input into the evaluation may be rationally sensitive to *evidence* about what you will do. This is the sort of decision-dependence that characterizes causal decision theory. CDT's evaluation in terms of causal expected utility uses as input the agent's credence in various causal situations. But as both Newcomb and Damascus cases show, our credence in causal situations can be rationally sensitive to evidence about what one will do.

The problems facing causal decision theory, I suggest, stem from its decision-dependent character. If an evaluation is decision-dependent, in either the metaphysical or the epistemic sense, then the possibility arises that each option fares worse, from the perspective of its own performance, than some other option. In this case, no action would be ratifiable. And so we can expect all decision-dependent views to face cases that are structurally like both the symmetric and asymmetric variants of Death in Damascus, with similar difficulties securing a plausible verdict and a stable justification. And any view that grounds what you subjectively ought to do in an evaluation that depends epistemically on your beliefs about what you will do will face objections about the illegitimate use of action-credences in deliberation.

Exploring the way these issues arise for other decision-dependent views in depth would be beyond the scope of a single paper. Nevertheless, I think appreciating this general diagnosis is important for a few reasons. First, we expect common problems to have common solutions. So there is some reason to favor a response to the problems plaguing CDT which generalizes to other decision-

[8] For a discussion of the decision-dependent character of prudential choices that change your future preferences, see Bykvist (2006) and Paul (2014). The connection between different decision-dependent views is explored in Hare & Hedden (2016).

dependent views, given that they face similar worries. Second, noticing that the problems for CDT also trouble a number of other independently attractive views should make us interested in a solution even if we would otherwise be satisfied to bite the bullet on behalf of either CDT or EDT. I think this motivation is especially important to bear in mind for those who are skeptical about heavy reliance on intuitions about the kinds of odd cases that we've been discussing; there is a deeper theoretical ambition behind the view we will develop.

I'll now propose and pursue a promising general thought about how to deal with these problems of decision-dependence. For reasons that will become clear, we will consider for the moment only choices with exactly two options.

III. A Promising Thought

One way of looking at the problems raised by decision-dependence is that we normally expect to be able to make a decision by looking at how options rank according to a single evaluation. But decision-dependence means that in certain choice situations, we are presented with two distinct *decision-relative evaluations* comparing options against each other – an evaluation from the perspective of one option being performed, and an evaluation from the perspective of the other option being performed. In cases of epistemic decision-dependence, these evaluations come from our beliefs conditional on performing each action, and in cases of metaphysical decision-dependence, they come from what the actual world would be like if those actions were performed.

Neither of these decision-relative evaluations seem to have any better claim, from the deliberating agent's point of view, than the other, and there is no guarantee that the evaluations will rank the options the same way. A view might try to force an evaluation from some privileged third perspective, perhaps one that in some way splits the difference between the aforementioned pair. In a sense, this is what the basic version of CDT does, taking the agent's initial action-credences as determining the relevant perspective. But we saw this strategy run into trouble – the agent's initial action-credences just don't seem to be the right thing to use as a privileged point of view in deliberation, and there aren't any other obvious candidates.

But what if instead of looking for a privileged perspective, we allow ourselves *both* evaluations, and settle what we ought to do based on the *relationship* between them? This is the core of the

promising thought. When our two evaluations agree our job is done – we pick the option that ranks highest. When they disagree, we must take a closer look.

We can represent the decision-relative evaluations for two-option cases in a matrix as follows:

	$\mathbf{V_A}$	$\mathbf{V_B}$
A	$V_A(A)$	$V_B(A)$
B	$V_A(B)$	$V_B(B)$

$V_A(A)$ and $V_A(B)$ give us the evaluation of the options from the perspective of A, and $V_B(A)$ and $V_B(B)$ give us the evaluation of the options from the perspective of B.

In our decision-theoretic example, the evaluation relative to X should be the causal expected utility of the options, from the perspective of X being chosen. We can capture this with the notion of *conditional causal expected utility* – effectively, what the causal expected utility would be after updating on the information that the agent chooses X. That is, $V_X(Y)$ is the causal expected utility of Y conditional on X, given by:

$$V_X(Y) = \sum_{i=1}^{n} \Pr(C_i \mid X)U(C_i \,\&\, Y).$$

To flesh this out, let us begin by applying it to *Newcomb's Problem*. Since we are trying to be good causalists, our decision-relative evaluations are determined by two calculations of the conditional causal expected utility of each option: first, conditional on one option being chosen, and second, conditional on the other.

In this case, we do end up with two different evaluations. Conditional on choosing one box, one-boxing is evaluated at 1000, and two-boxing is evaluated at 1001. Conditional on choosing two boxes, one-boxing is evaluated at 0, and two-boxing is evaluated at 1.

	$\mathbf{V_A}$	$\mathbf{V_B}$
One Box (A)	1000	0
Two Box (B)	1001	1

Of course, this looks exactly like the payoff matrix for Newcomb's problem we looked at earlier – this is because conditional on one-boxing, it is effectively certain that the box is full, and so

the conditional causal expected value of one-boxing is just the payoff for one-boxing if the box is full (with parallel reasoning for the other entries).

We notice that while the evaluations differ in an absolute sense, they agree about which option is better – two-boxing wins from the perspective of either action. So, naturally, we should think two-boxing wins overall.

Now let us contrast the symmetric (D) and asymmetric (D') versions of *Death in Damascus*.

D

	V_A	V_D
Stay in Aleppo (A)	0	1000
Go to Damascus (D)	1000	0

D'

	V'_A	V'_D
Stay in Aleppo (A)	100	1100
Go to Damascus (D)	1000	0

In both cases, from the perspective of staying in Aleppo, Damascus looks more attractive, and from the perspective of going to Damascus, Aleppo looks better. Our two evaluations will disagree. But each evaluation gives us not just an ordinal ranking of the options, but also the degree to which one option is better or worse than the other (the difference between the conditional causal expected utility of each option). In the symmetric case, the degree to which Damascus is more attractive, from Aleppo's perspective, is the same as the degree to which Aleppo is more attractive, from Damascus's perspective. In the asymmetric case, it is not the same. According to the promising thought, this comparison between the margins of victory in the two evaluations is precisely what explains why either option is permissible in the symmetric case, and only Aleppo is permissible in the asymmetric case.

In a sense, the thought is one expression of a commonsense explanation of why you should stay in Aleppo in *Asymmetric Death in Damascus* in terms of *regret*: You're going to regret what you decide no matter what, but you're going to regret it less if you stay in Aleppo.

More generally, in two-option cases where evaluations are decision-dependent, there will be three possibilities: a) one option ranks uniquely highest on both evaluations, b) no option ranks uniquely highest on both evaluations, and the margins of victory are equal, or c) no option ranks highest on both evaluations, and the margins of victory are lopsided. Our proposal is that if a) then

that option is uniquely permissible, if b) then both options are permissible, and if c) then the option with the greater margin of victory is uniquely permissible. Formally:

> **The Promising Thought:** For all actions X and Y, let $V_X(Y)$ be the decision-relative evaluation of Y from the perspective of X. For all two-option cases, an agent ought to perform X over Y iff $V_X(X) - V_X(Y) > V_Y(Y) - V_Y(X)$. [9]

This view is on track to give us everything we wanted. It two-boxes in *Newcomb's Problem*, and gets the intuitive results in the two-option cases that threatened CDT and its variants. Moreover, it gives no role to our current, unconditional credences in our actions – similar to EDT, we only look at the probabilities conditional on our actions. So it is immune to concerns about the illegitimate use of initial action-credences in deliberation.

There is, it is worth noting, a kind of case where the promising thought diverges from both CDT and EDT. Suppose you face a predictor who offers you a choice between box A and box B. You are also told the following: if you were predicted to choose A, the predictor put ten thousand dollars into box A. If you were predicted to choose B, the predictor put a million dollars into your bank account yesterday, and a thousand dollars into box B.

	Predict A	**Predict B**
A	10	1000
B	0	1001

[9] The promising thought lines up with Barnett's (ms.) 'graded ratifiability' account of pairwise preferability, though Barnett makes no attempt to turn it into a theory about what we ought to do, due to difficulties introduced by decisions with three or more options.

This is once again a decision-dependent case with no dominant option, and CDT once again has the implication that what you ought to do depends on what you currently think you will do. However, unlike the Damascus cases, EDT and the promising thought do not agree here. EDT tells us to pick box B, since conditional on choosing B, we expect to end up with over a million dollars, and conditional on choosing A, we expect to end up with only ten thousand. The promising thought tells us to pick box A, since conditional on choosing A, we expect to have ten thousand dollars *more than we would have had by picking B*, and conditional on choosing B, we only expect to have one thousand dollars more than we would have had by picking A.

Is this a problem for the promising thought? Bassett (2015), who discusses a case with this structure, takes both options to be intuitively rational. The promising thought does not say this,[10] but neither do EDT or CDT, so if this is the desired response, at the very least we are not at a disadvantage. CDT's answer, that it all depends on what we currently believe we will do, seems mistaken for the reasons we have already discussed. And it is hard to see a motivation for the view that A is the *only* permissible option that does not amount to the thought that we should act to give ourselves the best news about things outside our control, and therefore imply the wrong results in newcomblike cases. If we take the criticisms raised earlier seriously, then it is at best far from clear that the promising thought says anything unacceptable in these cases where it departs from both standard views.

IV. Three's a Crowd

Unfortunately, all good things come to an end, and it turns out to be not at all straightforward to generalize the promising thought to decisions with three or more options. Moreover, the most serious existing attempt to do so faces powerful objections of its own.

In two-option cases, the promising thought asked us to make a simple comparison of the direction and margin of defeat according to each of two evaluations: $V_A(A) - V_A(B)$ and $V_B(B) - V_B(A)$. Adding a third option C, however, not only adds another perspective, but a great deal of additional information about the relative evaluation of options according to each perspective. How to incorporate all of this in the spirit of the promising thought is not immediately obvious.

[10] It is actually possible to modify the promising thought slightly to accommodate this judgment. We might say that in cases where both options are ratifiable, both options are permissible, and it is only in cases where *neither* option is ratifiable that we need to compare margins of victory. This modified view is perfectly compatible with the rest of the picture we develop in this paper. Thanks to an anonymous reviewer for suggesting this variation.

The natural way to approach this problem would be to define an *overall* evaluation of the options as a function of the decision-relative evaluations, and instruct the agent to pick the option that ranks highest on this overall evaluation. What we'd like, then, is a general way to define an overall evaluation function O(X) in terms of the various $V_Y(Z)$ – one that agrees with the promising thought in two-option cases and gives reasonable verdicts in other cases.[11]

Sum Theory

Initially, one might be tempted to just take the sum of the decision-relative values of A from each perspective and compare them to the sum for each other option. That is, one might claim:

$$O(X) = \sum V_Y(X) \text{ for all } Y.$$

According to this *Sum Theory* (ST)[12], in the three option case the agent ought to pick A iff $V_A(A)+ V_B(A)+ V_C(A)$ is greater than both $V_A(B)+ V_B(B) + V_C(B)$ and $V_A(C) + V_B(C) + V_C(C)$. This is equivalent to the promising thought in the two option case. But it has absurd implications.

Brother in Babylon

Brother in Babylon is a case identical to *Asymmetric Death in Damascus* except that you have a third option – you can go to Babylon. Unfortunately, Death's more sadistic brother lives in Babylon. If you go to Babylon, you will be tortured and killed. If Death predicts you will go to Babylon (and if you go, he almost certainly will), he won't bother showing up there – his brother will do his work for him, after all. Instead, he'll vacation in Aleppo. The payoff matrix looks like this:

	Death predicts Aleppo	**Death predicts Damascus**	**Death predicts Babylon**
Stay in Aleppo	100	1100	100
Go to Damascus	1000	0	1000
Go to Babylon	-500	-500	-500

[11] This way of thinking about the task owes much to Briggs (2010), who proposes that we can understand decision rules as voting rules, where the options play the role of both candidates and voters. Here V_Y is analogous to a function describing voter Y's preferences among the candidates, and O provides a kind of aggregate rating on the basis of the individual ratings.

[12] As far as I am aware, nobody has defended this particular view, though it is a natural place to begin generalizing the promising thought.

Intuitively, you ought to stay in Aleppo. The addition of the ridiculous option of going to Babylon should not make any difference. But ST says otherwise.

As we saw earlier, because we have a practically infallible predictor, the entries in the payoff matrix conveniently do double duty as both utilities of outcomes and as giving the (approximate) decision-relative values of each choice (with the "Death Predicts Aleppo" column, for example, giving the values relative to staying in Aleppo.)

The overall value for Aleppo according to ST, then, is just the sum of the Aleppo row, or 100+1100+100 = 1300, while the overall value for Damascus is 0+1000+1000 = 2000. Babylon, which winds up dead last with a sorry -1500, turns out to be a spoiler for Aleppo, and in a bizarre sort of way – why should the performance of Aleppo relative to the completely absurd choice of going to Babylon bear on whether it is better or worse than Damascus?

These divergent recommendations in *Asymmetric Death in Damascus* and *Brother in Babylon* bring out a class of worries related to a principle called *Independence of Irrelevant Alternatives* (IIA). According to IIA, the addition of a new option should not affect the rankings of existing options relative to each other. As we will see later, I do not think IIA is sacrosanct – independent reasons to accept violations have been proposed in other domains.[13] But not all violations of IIA are equally implausible, and the violation here is particularly egregious in at least two ways. First, Babylon is not only a wrong choice – it is a *universally dominated* choice – in no possible causal situation is it better than any of the alternatives, and it is worse in some. Consequently, its causal expected utility is never greater, and sometimes less, than that of other choices, no matter how we conditionalize. Second, the introduction of Babylon changes the deontic status of other options from *prohibited* to *obligatory* and vice versa. This is more radical than violations of IIA that allow options to go from prohibited or obligatory to permissible, but not to jump past mere permissibility.

Benchmark Theory

A different way to try and generalize the promising thought leads to an approach we can call *Benchmark Theory* (BT). On this kind of view, a version of which is defended by Ralph Wedgwood (2013), we look at how an option performs, from its own perspective, relative to a kind of benchmark

[13] Kamm (2007), for example, discusses examples in deontological ethics.

determined by the overall performance of options from that perspective. The agent ought to choose the option that performs the best against its own benchmark. More formally, BT tells us:

$O(X) = V_X(X) - B_X$, where B_X is a benchmark determined by the X-relative evaluation of options.

Because the benchmark for X is determined only by the X-relative evaluation, BT avoids one of the strange features of Sum Theory – the overall ranking of two options will never be affected by their performance from some third perspective. But it also faces major problems.

We can first notice that there are different ways of setting the benchmark. We could set B_A to be the highest value of V_A. We could set it to be the lowest. We could set it to be the average.[14] In the two-option case it doesn't much matter where we set the benchmark – the options will rank the same regardless. But as Wedgwood shows, in three-option cases how you set the benchmark matters for what the theory recommends. One worry for Benchmark Theory, then, is that we do not seem to have any nonarbitrary grounds to select the benchmark one way rather than another.

Worse, however, it also entails violations of IIA at least as implausible as *Brother in Babylon*. These violations will look slightly different depending on how exactly we set the benchmark, but for illustration, let us assume the most straightforward elaboration of BT, which sets the benchmark for an evaluative perspective at the *average* value.

Sister from Babylon

Sister from Babylon is just like *Brother in Babylon* except that Death's helpful sister is on her way from Babylon to Aleppo. If Death predicts Aleppo, and you go to Babylon, she will catch you on the road and, since it is on her way, deliver you back into Death's hands. The payout matrix looks like this:

[14] We could even do something more complicated, if our evaluative theory is spelled out in more detail, like set B_a to be a weighted average of the value of the best outcome in each causal situation, or a weighted average of the value of the worst outcome in each causal situation. Wedgwood (2013) is committed to one of these more complicated treatments, because he builds his view atop a benchmark at the level of outcome value in each causal situation rather than a benchmark at the level of conditional causal expected value, as we do here. So, technically, for Wedgwood, what we call B_X is determined not by the values of $V_X(Y)$ but on some of the facts that go into *determining* $V_X(Y)$. Wedgwood has theoretical reasons for developing the view in this way, but for the purposes of this paper, these more complicated treatments have analogous implications, and our way of framing the issue more easily generalizes to decision-dependent views outside this dispute in decision theory. Moreover, conveniently, taking the benchmark in his sense as the average value of the options in a causal situation, which he endorses, is equivalent to taking the benchmark in our sense as the average conditional causal expected value, making the views mere notational variants in the special case of averaging.

	Death predicts Aleppo	**Death predicts Damascus**	**Death predicts Babylon**
Stay in Aleppo	100	1100	100
Go to Damascus	1000	0	1000
Go to Babylon	100	-500	-500

The benchmark for staying in Aleppo is the average causal expected utility of the options conditional on staying in Aleppo, or (100+1000+100)/3 = 400. Staying in Aleppo falls short of this benchmark by 300. The benchmark for going to Damascus is (1100-500)/3, or 200. Going to Damascus falls short of this benchmark by 200. The benchmark for going to Babylon is (100+1000-500)/3 = 200, and Babylon falls short by 700. So you ought to go to Damascus. Once again, the introduction of Babylon, a third, universally dominated option, has turned the recommendation from *Asymmetric Death in Damascus* on its head. So BT also entails severe violations of IIA.

It is bad enough that dominated options sometimes affect the evaluations of other options. But it gets even worse. Sometimes, BT actually entails that dominated options are *obligatory*.

Reunion in Babylon

In *Reunion in Babylon*, the rest of Death's extended family, spread out across every city, is planning to hold a reunion in Babylon. If Death decides to go to Babylon, then one of them will notice you and deliver you to Death, no matter where you are. If Death goes to Aleppo or Damascus and can't find you, he'll call his family in Babylon and they will do the dirty deed for him if you are there. And you find the prospect of dying in Babylon particularly unpleasant.[15]

	Death predicts Aleppo	**Death predicts Damascus**	**Death predicts Babylon**
Stay in Aleppo	100	1100	-1
Go to Damascus	1000	0	-1
Go to Babylon	-1	-1	-1

[15] This case has the structure of the "Three Option Smoking Lesion" discussed by Egan (2007), attributed to Anil Gupta.

Again, going to Babylon is universally dominated – it can only do worse than the other options. But while both Aleppo and Damascus significantly underperform their benchmarks, Babylon does not underperform its own at all. So BT tells us that *Babylon* is the correct choice. This is a disaster – after all, this view was motivated by trying to preserve two-boxing in Newcomb's problem, and the case for two-boxing is a causal dominance argument. Yet the view is now demanding that we choose a causally dominated option.

Wedgwood is aware of these objections. While he does not accept IIA, he agrees that dominated options should not make a difference, and certainly should not themselves be chosen. His solution is to add to BT a requirement that, before even calculating the benchmark, we exclude all options that "do not deserve to be taken seriously" from consideration – in particular, dominated options. But this is a deeply unsatisfying move for a number of reasons. It looks unacceptably ad hoc – EDT would also get the right answer in Newcomb's problem if we introduced such a requirement. And it gets things backwards – the reason that it's fine to ignore obviously insane choices in deliberation is precisely that we think including them would not make any difference and that a reasonable procedure would never lead us to choose them. We should not exclude them out of fear that if we didn't, our decision procedure would render them victorious. Finally, Briggs (2010) points out that this move can be finessed just by altering the counterexamples slightly, making the relevant option *nearly dominated* rather than dominated. For example, we could alter the payoffs in *Reunion in Babylon* as follows:

Nearly-Dominated Reunion in Babylon

	Death predicts Aleppo	**Death predicts Damascus**	**Death predicts Babylon**
Stay in Aleppo	100	1100	-1
Go to Damascus	1000	0	-1
Go to Babylon	-1	-1	-0.99999999999

Now, Babylon is no longer dominated. And it wins, handily, being the only option that doesn't massively underperform its own benchmark. But if the payoff for running into death in Babylon was even a mosquito-bite worse, it would be dominated, and therefore excluded. This is a bizarre discontinuity – adding this tiny payoff should not turn Babylon from an option so insane that it should

be excluded from consideration outright to an option so much better than the alternatives that choosing it is required. Similarly, if violations of IIA arising from the introduction of dominated options are unacceptable, violations arising from the introduction of nearly-dominated options don't seem much better.

The Self-Esteem Theorem

One way to characterize the difference between ST and BT is to notice that in our examples, ST calculates the overall value of an option by looking at the relevant *row* of the payoff/evaluation matrix – the overall value of Aleppo is determined by the values in the "Stay in Aleppo" row. BT, by contrast, calculates the overall value of an option by looking at the relevant *column* of the matrix – the overall value of Aleppo is determined by the values in the "Death predicts Aleppo" column. Both of these ran into problems with IIA – Babylon's presence in Aleppo's row allowed it to be a spoiler for Aleppo under ST, and Babylon's presence in Aleppo's column allowed it to be a spoiler for Aleppo under BT.

But perhaps, one might hope, we just haven't been creative enough with our method of determining overall value from the decision-relative evaluations. Unfortunately, we can prove a theorem that will all but stamp out this hope.

Take a decision-relative value matrix for three-option decisions, and let O(X) be the overall evaluation of X, as a function of the various $V_Y(Z)$:

	$\mathbf{V_A}$	$\mathbf{V_B}$	$\mathbf{V_C}$
A	$V_A(A)$	$V_B(A)$	$V_C(A)$
B	$V_A(B)$	$V_B(B)$	$V_C(B)$
C	$V_A(C)$	$V_B(C)$	$V_C(C)$

Suppose, noticing the counterexamples to ST and BT, we want our overall evaluation to respect IIA. The addition of a third option, in other words, should not change the overall comparison between existing options. What does this mean exactly? Let P(X, Y) be a function representing this overall comparison, for instance, one that assigns 1 if O(X)>O(Y), 0 if O(X)=O(Y), and -1 if O(X)<O(Y). What IIA tells us, then, is:

*IIA***:** P(X, Y) is a function P' of $V_X(X)$, $V_Y(Y)$, $V_X(Y)$, and $V_Y(X)$.

This is to say that none of the entries in the C row and the V_C column should make a difference to whether O(A)>O(B). Note that this is a weak assumption – we are not assuming that the C-related entries cannot make a difference to O(A) or O(B) themselves, nor even that they cannot make a difference to the cardinal value of O(A)-O(B). But even with this weak constraint, it turns out that there is only one way for this to hold.

> *The Self-Esteem Theorem*: Given IIA, P(X, Y) is a function of X and Y's *self-esteem* – that is, $V_X(X)$ and $V_Y(Y)$ alone. [The proof of this theorem is given in the appendix]

The Self-Esteem Theorem is bad news for the project of generalizing the promising thought, which denies that in two-option cases you ought to pick the option with the higher self-esteem. In fact, recall that in our application, X's self-esteem $V_X(X)$ is the causal expected utility of X conditional on X being performed – in other words, its evidential expected utility. This means that the only view that preserves overall comparisons with the addition of new options and can be cast in terms of decision-relative value is our old enemy EDT.[16] And we know EDT one-boxes in Newcomb's problem, which is exactly what we wanted to avoid.[17]

Taking Stock, Part II

If the promising thought is the hero of this paper, then we are now at the point in the narrative where they have hit their lowest point. The villain seems beyond defeat – every plan by our hero has failed. The promising thought was attractive because it seemed like it could give us our desired result in Newcomb's problem without the costs of CDT. And for a while, it did. But as soon as a third option was introduced, we ran into problems. Every attempt at generalization ran into massively counterintuitive violations of IIA, and sometimes worse. Moreover, we now have arguments to the effect that *any* way to make sense of decision-dependent evaluations will be forced to violate IIA or

[16] The only *reasonable* view, at least. Technically, any overall value function in which O(X) is a function of EEU(X), even the trivial one that always assigns every option the same value, will also pass.

[17] A different but related result can be found in Briggs (2010). Translated into the general framework we are using for decision-dependent evaluation here, Briggs effectively shows that no overall evaluation can satisfy both of the following principles: **P:** For any actions A and B, if $V_X(A) \geq V_X(B)$ for all X, and $V_X(A) > V_X(B)$ for some X, then O(A) > O(B). **S:** Let V and V' be the decision-relative evaluation functions for two decision problems D and D', and O and O' be the overall evaluation functions for those two decision problems. For any actions A and B, if $V_A(X) = V'_a(X)$ for all X, and $V_B(X) = V'_B(X)$ for all X, then O(A) > O(B) iff O'(A) > O'(B). The first is a kind of *pareto-optimality* constraint. The second is a kind of *self-sovereignty* principle slightly weaker than the IIA we use above.

abandon the promising thought. For two-boxers in Newcomb's problem, and for anyone who finds decision-dependent views appealing, things look grim indeed.

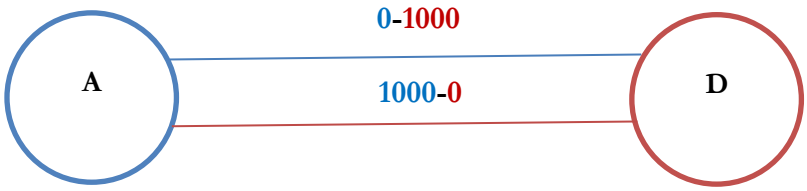
Things look so grim, in fact, that I think a diversion would be good for our psychological health. I promise the reader that if they are patient, they will get an ending with a happier fate for our intrepid protagonist. Meanwhile, let's interrupt our dry and demoralizing discussion of decision theory to tell a story about my favorite fictional sport, Calvinball.

V. Intermission: A Brief History of Calvinball

Calvinball is always played between exactly two teams, who each try to score as many points as possible. Though it began as a casual diversion, soon enough the International Calvinball League was formed to settle the question of which team was the best.

The League quickly ran into a problem. Calvinball developed largely independently in many different places at the same time, and there was no common standard for the geography of the playing field. Each team had its own distinctive home field, and there was no obvious way to pick a privileged place for two teams to hold their match. So the League ruled that each match would be composed of two individual games, one on each team's home field, and the scores would be combined.

Here is a representation of a notoriously controversial match in the Mesopotamian division, between Aleppo and Damascus:



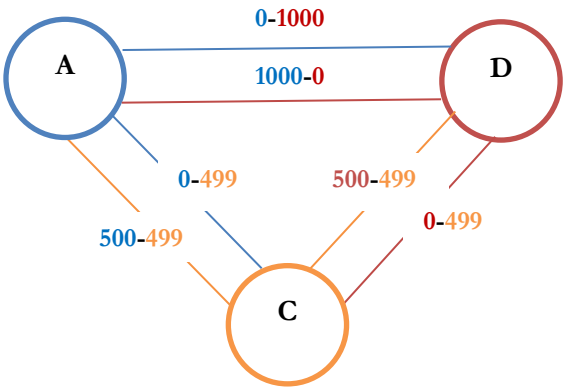
Each colored circle is a team, and each line is a game played between them, the red line representing a game played on the red team's field, and the blue line a game played on the blue team's field. The score for each game is above the line.

This outcome was something of an embarrassment for everyone involved – each team lost on their own field. And the scores were symmetrical, so there was nothing for the officials to do but declare the match a draw. Nobody was happy with this result, so after a period of intense training, a hotly anticipated rematch was held:

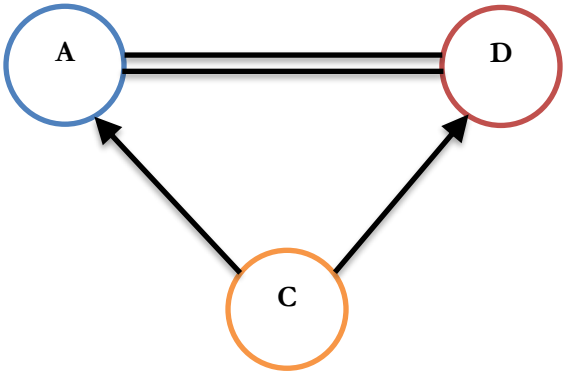
The teams split games again, but this time the result was not symmetrical – Aleppo's margin of victory was significantly greater, and was declared the victor.

So far this has been an account of how individual matches of Calvinball are scored. Of course, there are often more than two teams in a Calvinball competition. For a larger tournament, the League decided to choose a round-robin format, where each team plays amatch against each other team.

To illustrate, here is a tournament in which the Aleppo and Damascus teams competed with a new upstart team, the Coinflippers:

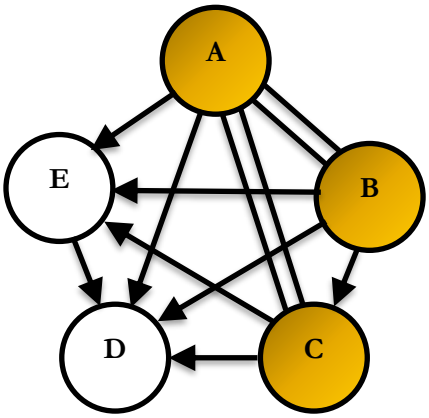


Once again, each team split individual games against each other team. But because of the margins of victory, the Coinflippers won both of their *matches*, while Damascus and Aleppo each tied once and lost once. For the purposes of scoring the tournament, the League decided to take into account only information about which teams defeated which, and we can simplify our diagram to reflect this. The arrows point from the winners to the losers of individual matches, and parallel lines represent a draw. We can call this the *defeat graph* for the tournament:



Now, we haven't said anything yet about exactly how the overall winner or winners are chosen based on a tournament's defeat graph, but here it is not hard to guess; any reasonable way of scoring, it was decided, would respect the *Condorcet Principle*: if a team defeats every other team, then it wins.

Past this, however, there was some disagreement over how to score the overall tournament. One tournament with five teams ended up like this:



The controversy over the winner was heated. B argued that they alone should win, and A argued that it should share the victory with B, but ultimately the League opted for a conservative view. They decided that a winner had to defeat *all* of the losers. Here, A, B, and C each beat D and E. But none of A, B, and C beat the other two. So the rules demanded that A, B, and C had to share the crown, represented here by their nodes being filled in with champion's gold. There was some grumbling, but the League had a method for picking a winner or winners in every possible Calvinball tournament, and so it has remained ever since.

VI. Tournament Decision Theory

As the reader will have surmised, my proposal is that we understand decision problems by analogy with the story I just told. According to the view I will call *Tournament Decision Theory* (TDT), we can understand a choice situation as a tournament between options. These options do not all compete against each other on the same field at the same time – each option battles each other option separately, head-to-head. The deontic status of each option is determined by the results of these matches.

TDT is less a specific view than a template for building a view; there will be many incompatible views which fall under the TDT framework. We can think of each possible elaboration of TDT as having two essential parts: first, it has a way to determine the head-to-head *match results* between options – each match should have a winner and a loser, or be a tie. Second, we need a way to pick the winners of the tournament based on the match results – that is, based on the tournament's defeat graph. A function from a tournament's defeat graph to the subset of the players representing the winners is what mathematicians and social choice theorists call a *tournament solution*.[18] Tournaments have already been studied mathematically in depth as potential solutions to voting problems; here we will borrow this tool as a way of dealing with individual decision-making in decision-dependent circumstances.[19]

Ultimately, what is most important to us in the context of decision-making is placing our options in deontic categories – dividing our options into the permissible and the impermissible. The "winners" of the tournament, then, are the permissible options. The other options are impermissible. A fully developed Tournament Decision Theory could have more fine-grained tournament rankings to capture the sense in which some impermissible options are worse than others, but for now we will set that aside.

So far, this framework is extremely general. It is general enough to accommodate CDT and EDT – if we determine match results by comparing evidential expected utility or causal expected utility, then any reasonable tournament solution will deliver the results we expect from those views.

But this is not a particularly illuminating way to understand those approaches. That is because both CDT and EDT compare any two options according to their place on a single fixed evaluation no matter which two options we are considering – the evaluation in terms of CEU or EEU respectively. In this way, CDT and EDT treat the competition between options as rather like weightlifting. There is a certain amount that each weightlifter can lift, and it doesn't (we'll assume) vary across contexts or depend on who they are lifting against. You *could* run a weightlifting competition by having each weightlifter square off against each other weightlifter in a head-to-head

[18] Because we allow ties between options, we are dealing with what are sometimes called "weak tournaments", in contrast to "strong tournaments" which require that every match result have a winner and loser. Most mathematical work on tournaments concerns strong tournaments. Fortunately, there are systematic ways to generalize common tournament solutions to weak tournaments while preserving many of their attractive properties. See Brandt, Brill, & Harrenstein (2014).
[19] Laslier (1997) summarizes roughly a decade of early work on tournament solutions in social choice.

match, but this would be wasteful – you'll get the same results more straightforwardly by just checking once how much each weightlifter can lift and then giving the crown to whoever lifts the most.

TDT starts to become interesting as a genuine alternative when we do not have this kind of single fixed evaluation, and in particular when it makes sense to compare options two at a time – that is, when the competition between options seems more analogous to Calvinball than weightlifting. It is especially appealing when our method of comparison looks *essentially contrastive*. Thinking about decisions in terms of *regret* is one example. Arguably, regret is always relative to some particular other option – one regrets doing this *rather than* that. It is also appealing when we have a method of comparing two options that does not easily generalize to three options or more.

This should be sounding familiar, of course. These are exactly the conditions we found ourselves in trying to make sense of decision-dependence in general, and in our attempt to avoid the pitfalls of CDT in particular. I promised that the promising thought would have a happier ending. Tournament Decision Theory is the help it needs.

The Promising Thought Redux

For the rest of this paper, I will explain how a fleshed-out version of TDT might handle cases of the sort we have been discussing. The promising thought is ready to take the spotlight again – this time as the way to settle individual matches between options. That is, for any two options X and Y:

i) If $V_X(X) - V_X(Y) + V_Y(X) - V_Y(Y) > 0$, X defeats Y.
ii) If $V_X(X) - V_X(Y) + V_Y(X) - V_Y(Y) = 0$, X ties with Y.

This is enough to determine the results of any match between two options. Before we have any deontic implications, however, we have to add a tournament solution. There are a number of interesting prima facie plausible candidates for such solutions, each of which will have different properties. For example:

1) The *Top Cycle* (or Smith Set): the smallest set of options that defeats every option outside the set.

2) The *Uncovered Set*: the set of all options that are not *covered* by other options, where X covers Y if anything that defeats X defeats Y, and either something defeats Y that doesn't defeat X or X defeats something that Y doesn't defeat.[20]
3) The *Banks Set*: the set of all the top options in maximally extensive transitive sub-tournaments.
4) The *Bipartisan Set*: the set of all options that are played with some probability in the Nash equilibrium strategy of a game in which each player selects an option, and wins if it defeats the opponent's.

Assessing the relative virtues of these and other possible tournament solutions is an important task for a complete development of tournament decision theory, but it is beyond the scope of this paper.[21] They share certain general properties in common – for instance, they are all *Condorcet* solutions, in virtue of the fact that if an option beats every other option, it is always selected as a unique winner. In small tournaments, they overlap significantly. In fact, in all the cases we've discussed, their implications are identical.

For illustration, however, I'll assume the *Top Cycle* as our tournament solution, as the International Calvinball League did in our story. This is not because the top cycle is unambiguously the best solution; in fact, it is generally thought to be too conservative in picking winners. It is the least *discriminating* of the tournament solutions that are seriously considered – every plausible alternative picks a subset of the top cycle as winner. But it is discriminating enough to get results in the cases we've worried about, and it will be useful as an example of how a tournament solution can place meaningful constraints on potential counterexamples.

With this in hand, I give the defeat graphs for each of the decision problems we have discussed so far:

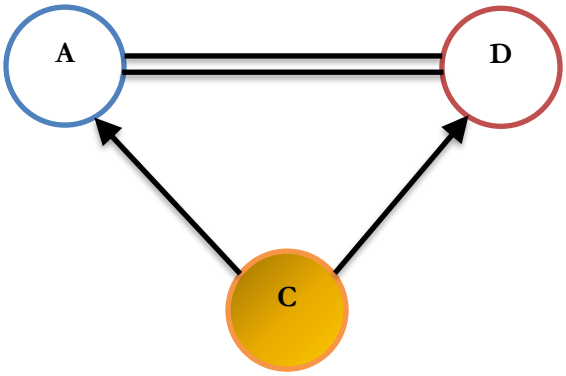
[20] This is one interpretation of a covering relation. See Penn (2006) for discussion of alternative covering relations and their properties.
[21] For an overview of the common proposals solutions see Laslier (1997).

Newcomb's Problem

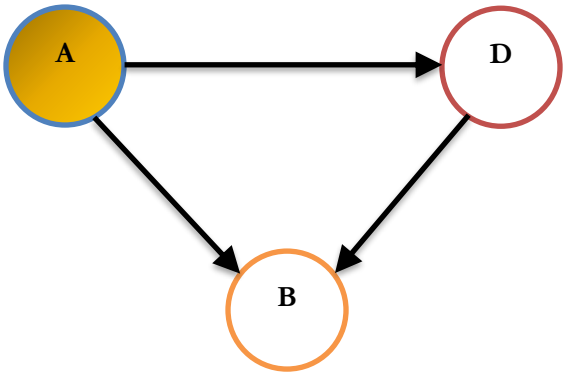
Death in Damascus

Asymmetric Death in Damascus

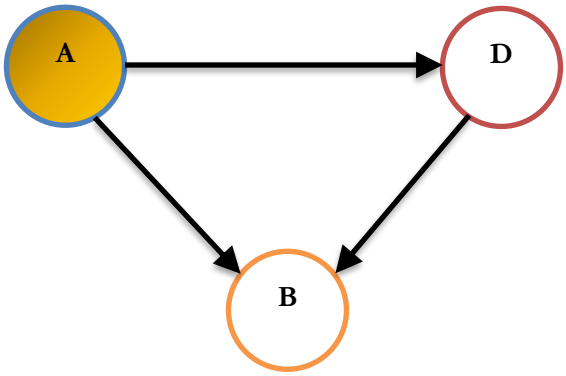
Dicing with Death



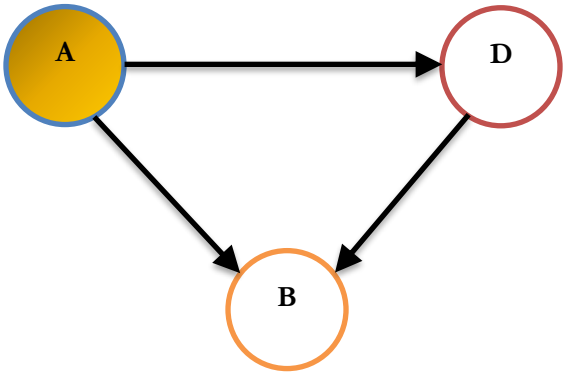
Brother in Babylon



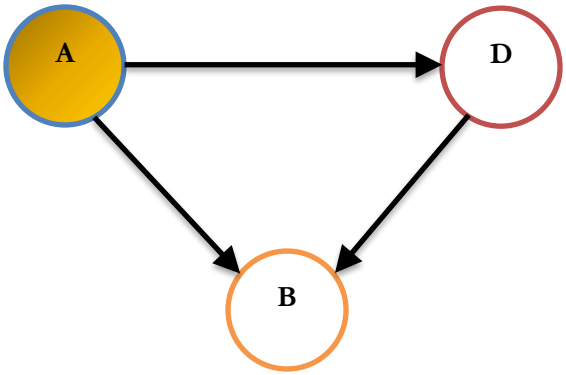
Sister from Babylon



Reunion in Babylon



Nearly-Dominated Reunion in Babylon



This looks encouraging! All of these results are intuitively correct. So TDT has been able to get an attractive set of implications that have eluded all our other attempts. Because match results are determined in accordance with the promising thought, and the winner of a two-option tournament is just the winner of the match between the options, TDT is trivially going to agree with the promising thought in all such cases. But it does not fall to the counterexamples against other extensions of this thought. [22]

For all we've said so far, however, it might be that we simply haven't chosen the right examples. Didn't the Self-Esteem Theorem show us, after all, that no view that respects the promising thought was going avoid violating IIA, and doesn't that mean that some counterexamples must be lurking of the sort that sink the Sum and Benchmark theories?

For the rest of the paper, we will look carefully at the relationship between tournament decision theory, dominance, and the independence of irrelevant alternatives, and we will see why the nature of our tournament solution guarantees that we cannot construct counterexamples as severe as those facing other extensions of the promising thought.

Not All Irrelevant Alternatives Are Equally Irrelevant

Let us return to the Self-Esteem Theorem. What that theorem showed is that there is no way to assign an overall value to options on the basis of their decision-relative evaluations that both a) preserves the overall comparison between two options with the addition of a third, and b) doesn't ground the overall evaluation purely in each option's self-esteem. What does this mean for TDT?

The first thing to note is that there is a clear sense, for the tournament decision theorist, in which comparison between options is unaffected by additional options. The individual *match* results are a kind of comparison, and they are determined entirely by four decision-relative values which are not affected by further additions to the option set.

The second thing to note is that unlike the Sum and Benchmark theories, TDT does not determine an option's deontic status on the basis of its place on any assignment of overall value to each option – it does it on the basis of essentially pairwise comparisons. These pairwise comparisons

[22] The reader may have noticed that there are two possible defeat graphs with three options which we have not considered. They do not correspond to any of the cases we have been discussing, and different plausible tournament solutions give different verdicts about cases represented by those graphs, so we may set them aside for our purposes.

do not correspond to relative placement on any single evaluative scale. So there isn't any overall evaluation, in the same sense, for which we can even *ask* whether IIA holds.

Now, we *could* take the assignment of deontic status itself as a kind of overall evaluation – for example, setting the overall value of an option as 1 if TDT judges it permissible and 0 otherwise. The third thing to note is that in fact *nobody* accepts IIA as we've formulated it applied to deontic status in this way. Everyone agrees that adding a third option can affect the relative permissibility of existing options in at least the following way: if between A and B, A is permissible and B is not, then the addition of a third option can change the relative permissibility of A and B if C is *obligatory*. Of course the addition of a third option can change the relative permissibility of A and B if A is the better of the two, but only the second best of the three.

It would be nice to say that this kind of "violation" of IIA is the only sort licensed by TDT. But it does license another kind of violation which not every theory accepts. This is guaranteed by the non-transitivity of defeat. The following is a case that Hare & Hedden (2016) use as part of an argument against CDT:

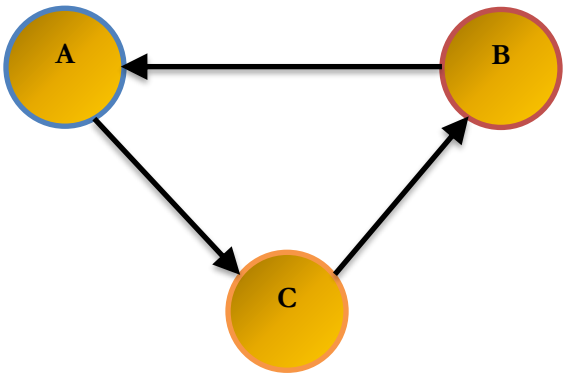
Three Crates

A perfectly reliable demon has made a prediction about which of three boxes you will choose, and filled them appropriately as follows:

	Demon Predicts A	**Demon Predicts B**	**Demon Predicts C**
Choose A	1000	0	0
Choose B	1001	0	0
Choose C	0	1	0

The causalist, Hare & Hedden claim, must say something quite implausible: that C is the only permissible option. Their argument, roughly, is that a rational, self-aware adherent of CDT will not choose an unratifiable option.[23] Choosing A is not ratifiable, because conditional on Choosing A, Choosing B looks better by CDT's lights. Choosing B is not ratifiable, because conditional on Choosing B, Choosing C looks better by CDT's lights. But conditional on choosing C, choosing C is fine, because the prospect of everything is equally dim. And the reader can check for themselves that indeed, on the three variations of CDT that attempt to respect ratifiability, choosing C is the only potentially permissible option.

What does TDT say? Here is the defeat graph:



A, B, and C are in a *defeat cycle*. On any tournament solution, this means the tournament is a three-way tie, and all options are permissible. We may notice, however, that between A and B, we have a case of causal dominance. So A would be impermissible if those were the only two options. This is a case where TDT entails that adding a third option can take an option that was previously *impermissible*, and render it permissible. And while every theory allows that a third option can kick a permissible option down into impermissibility, not every theory allows that a third option can kick an impermissible option *up* into permissibility.

This is one additional violation of IIA that TDT must accept. But not all violations of IIA are equal, and I think this one is relatively plausible. I think it is not absurd to respond to *Three Crates* with something like the following thought: "I should pick B over A, because B dominates A. But between B and C, I should pick C, since it dominates *in the situations I expect to be in*, conditional on picking either.

[23] Actually, the principle they use is a bit weaker than this, but the details don't matter for our purposes here.

And between C and A, I should pick A, for the same reason. So no option seems unequivocally better than the others, and I might as well pick any." Or in terms of regret: "As soon as I've made my choice, I'll regret choosing A over B, but not vice versa. I'll also regret choosing B over C. And I'll regret choosing C over A. So no option seems unequivocally better than the others, and I might as well pick any."

Three Crates, in other words, is a genuinely puzzling case, and while I think Hare and Hedden are right that picking C is not obligatory, TDT's shrug of a judgment does not look patently unreasonable, even though it entails bumping up an option's permissibility from some two-option Newcomblike cases. At the very least, it looks like an *improvement* on the judgment of CDT. Even if the reader is not completely sold, I want to at least make the case that this is about as bad as it gets for TDT, and contrast it to what I think are the much more serious problems faced by the Benchmark and Sum Theories.

Dominance and the Influence of Alternatives

Let us say an option *strongly DR-Dominates* another option if it is better on every decision-relative evaluation. In our decision-theoretic application, any option that is strongly causally dominated is also strongly DR-dominated, and any option which is weakly causally dominated is strongly DR-dominated on the minimal assumption that the agent's conditional credence in each causal situation is non-zero. So what goes for DR-dominance goes for causal dominance as well; for convenience, I will just refer to strong DR-dominance as *dominance*. The *Top Cycle*, recall, chooses as permissible the smallest set that defeats all options outside the set. Here are a few more or less straightforward facts about defeat, dominance, and the top cycle.

1) If X dominates Y, then X defeats Y.
2) If X defeats Y, then Y is in the top cycle only if X is.
3) If all options defeat X, then X is not in the top cycle.
4) If X is in the top cycle, then X remains in the top cycle with the addition of Y unless Y defeats all options.
5) If X is not in the top cycle, then X remains outside the top cycle with the addition of Y unless Y is in the top cycle.

6) If X defeats all options, then X is the only member of the top cycle.[24]

From this, several things follow about deontic status:

7) (From 1, 2) No dominated option is ever permissible unless the option that dominates it is also permissible.
8) (From 7) No dominated option is ever obligatory.
9) (From 1, 3) No option that is dominated by every other option is permissible.
10) (From 4, 5) Adding an impermissible option never affects which options are permissible.
11) (From 1, 2, 4) Adding a dominated option never makes a permissible option impermissible.
12) (From 6) If an option dominates all other options, then it is obligatory.

Evidential Decision Theory notoriously violates 12 in Newcomb cases. Benchmark Theory, without the ad hoc stipulation that dominated options be removed from consideration from the outset, violates all but 12. Sum Theory violates 10 and 11. The *Babylon* cases are so counterintuitive, I suggest, precisely *because* they are violations of these plausible constraints. Respect for these constraints, however, simply falls out of the properties of our tournament solution, and effectively guarantees that cases like *Three Crates*, where adding a permissible option bumps an impermissible option into the winner's bracket, is the only suspicious violation of IIA we'll be able to find.

Like Wedgwood, then, we accept that our view violates IIA in some respects. But we are able to systematically preclude the most implausible violations without resorting to undermotivated outright exclusion from consideration. The way TDT handles this is also superior to Wedgwood's strategy in another respect. Excluding dominated options creates a kind of *discontinuity* around the exact point at which an option becomes dominated. A little below, and the option doesn't get to participate at all; a little above, and the option suddenly is able to interfere in all the ways the dominated option would have. So tiny changes in the utility of a single action in a single outcome can lead to quite radical alterations in the way options are evaluated. But for TDT, a small change in an evaluation brought about by a small modification to an outcome can only affect the permissibility of options if it affects the defeat relations, and it only affects the defeat relations if the match between two relevant options was already very close. So we won't have counterexamples like *Nearly-Dominated Reunion in*

[24] All of these properties of the top cycle are trivial to prove and well-understood. Laslier (1997) contains discussion and proofs of these and other properties of the top cycle and other tournament solutions.

Babylon where an option's influence over a decision problem is subject to a drastic shift around the point of dominance.

Conclusion

What I have tried to develop in this paper is an approach to decisionmaking which, though available in principle any theorist, should be particularly appealing to those attracted to views on which evaluation is decision-dependent. I believe it is the best generalization of what I called the promising thought – a natural idea about how to deal with decision-dependence in two-option cases. Applied to the case of causalist decision theory, it opens the door to a view which has intuitive advantages over standard causal views, and over existing attempts to reconcile the causalist's answer to Newcomb's Problem and the evidentialist's answer to cases like Asymmetric Death in Damascus. Moreover, it is not vulnerable to theoretical worries about the use of initial action-credences in deliberation.

Even if one is convinced for independent reasons that the evidentialist was right all along, or that the virtues of causal decision theory are sufficient to outweigh the counterexamples, there are many other views which are decision-dependent, and we can expect strong parallels between the problems they face. Cases with the structure of *Three Crates*, for instance, are presented by Hare & Hedden as an objection not just to causal decision theory, but to views about the moral importance of children and about the relevance of future desires. The tournament approach to resolving these problems offers new opportunities for these views as well – it may turn out that the most fruitful application of our framework is outside decision theory altogether.

Finally, I think we can also see a path forward for further development of Tournament Decision Theory. We have shown how one specific tournament solution constraints the set of permissible options in ways that guarantee certain particularly worrying counterexamples cannot be constructed. But this should be seen as kind of proof of concept; there are many possible tournament solutions, with a range of different properties, and I have not argued that *Top Cycle* is the best. Indeed, we should not assume that the best tournament solution in one domain is necessarily the best in all of them. Looking to the mathematics and social choice literature that already exists on tournament solutions can help us see the advantages and disadvantages of different variants on tournament decision theory and anticipate the shape of potential objections. Notably, while the tournament view is flexible, it is not unlimitedly so – as we mentioned, every plausible tournament solution is a subset of the top cycle. And counterexamples to the tournament approach to a domain could be found, if,

for example, there are cases which share a defeat graph but about which our intuitions strongly differ. In the end, Tournament Decision Theory represents a significant departure from the standard way to look at deliberation, and it may not wind up the winner in all the fields in which it might compete. But I hope I've made the case that it should be taken seriously as a contender.

Appendix: Proof of the Self-Esteem Theorem

Proof of the Self-Esteem Theorem

Let O(X) be the overall evaluation of X as a function of the decision-relative values $V_Y(Z)$. Let P(X, Y) be a function representing the relative overall evaluation of X and Y, such that P(X, Y)=1 if O(X)>O(Y), 0 if O(X)=O(Y), and -1 if O(X)<O(Y).

IIA: P(X, Y) is a function P' of $V_X(X)$, $V_Y(Y)$, $V_X(Y)$, and $V_Y(X)$.

Proof:

Consider a problem with the following decision-relative value matrix:

	$\mathbf{V_A}$	$\mathbf{V_B}$	$\mathbf{V_C}$
A	p	q	p
B	r	s	t
C	p	u	p

By IIA, P(A, C) = P'(p, p, p, p) = P(C, A). It follows from the definition of P that, for any X and Y, P(X, Y) = -P(Y, X). So, P(A, C) = -P(A, C), which entails:

1) P(A, C) = 0

By the definition of P, this means that O(A)=O(C). Subtracting O(B) from both sides of the equality, we get:

2) O(A)-O(B) = O(C)-O(B)

3) P(A, B) = P(C, B) [Follows from 2 and the definition of P]

4) P'(p, q, r, s) = P'(p, q, t, u) [Follows from 4 and the definition of P' in terms of P]

5) So, the Self-Esteem Theorem is true: P is a function of $V_X(X)$ and $V_Y(Y)$ alone.

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

Barnett, David James. (ms.) "Graded Ratifiability."

Bassett, Robert. (2015) "A Critique of Benchmark Theory." *Synthese* 192(1): 241-267.

Brandt, Felix, Brill, Markus, and Harrenstein, Paul (2014). "Extending Tournament Solutions." *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*: 580-586

Briggs, Ray (2010). "Decision-Theoretic Paradoxes as Voting Paradoxes." *Philosophical Review* 119 (1):1-30

Bykvist, Krister (2006). "Prudence for Changing Selves." *Utilitas* 18 (3):264-283

Gibbard & Harper (1976). "Counterfactuals and Two Kinds of Expected Utility." in *Ifs*, Harper, Stalnaker, and Pearce (eds.), Reidel, Dordrecht.

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

Hajek, Alan (2003). "What Conditional Probability Could Not Be" *Synthese* 137:273-323

Hajek, Alan (2016). "Deliberation Welcomes Prediction." *Episteme* 13 (4):507-528

Hare, Caspar and Hedden, Brian (2016). "Self-reinforcing and Self-frustrating Decisions." *Noûs* 50 (3):604-628

Joyce, James (2002). "Levi on Causal Decision Theory and the Possibility of Predicting One's Own Actions", *Philosophical Studies* 110 (1): 69 – 102.

Joyce, James (2012). "Regret and Instability in Causal Decision Theory." *Synthese* 187 (1):123-145

Kamm, Frances (2007). *Intricate Ethics*. Oxford University Press, Oxford.

Laslier, J-Francois (1997). *Tournament Solutions and Majority Voting*. Springer-Verlag Berlin Heidelberg, Berlin.

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

Levi, Isaac (1993). "Rationality, Prediction, and Autonomous Choice." *Canadian Journal of Philosophy* 23 (1), 339-363.

Louise, Jennie (2009). "I Won't Do It! Self-Prediction, Moral Obligation and Moral Deliberation" *Philosophical Studies* 146 (3):327-348

Paul, L.A. (2014). *Transformative Experience.* Oxford University Press, Oxford.

Penn, Elizabeth (2006). "Alternate Definitions of the Uncovered Set and their Implications." *Social Choice and Welfare* 27(1):83-87

Price, Huw (1991). "Agency and probabilistic causality." *British Journal for the Philosophy of Science* 42 (2):157-176

Price, Huw (2007). "The Effective Indexical", http://philsciarchive.pitt.edu/4487/.

Rabinowicz, Wlodek (2002). "Does practical deliberation crowd out self-prediction?" *Erkenntnis* 57 (1):91-122

Spohn, Wolfgang (1977). "Where Luce and Krantz Do Really Generalize Savage's Decision Model." *Erkenntni*s 11 (1) 113–134

Spohn, Wolfgang (1978). *Grundlagen der Entscheidungstheorie*. Scriptor, Kronberg/Ts

Velleman, David (1999). *The Possibility of Practical Reason.* Oxford University Press, Oxford.

Wedgwood, Ralph (2013). "Gandalf's Solution to the Newcomb Problem." *Synthese* 190 (14):2643–2675