Prediction Impairs Deliberation

August 2023

Abstract

Casual decision theory has typically been discussed in contexts where agents adopt sharp credences of the sort adequately represented by a single probability measure. Under this assumption, causal decision theory is committed to a number of counterintuitive verdicts in cases involving decision instability. I suggest that many of these troubles lessen for sufficiently open-minded causalists who adopt maximally imprecise credences regarding their own acts. This provides novel pragmatic support for a version of the deliberation crowds out prediction thesis according to which rational agents' act credences should generally be imprecise in this way.

1 Introduction

Causal decision theory (CDT) has fallen on hard times. Once the dominant account of rational choice among philosophers, recent years have subjected CDT to a steady stream of criticism inaugurated by renewed concern regarding the theory's performance in contexts of *decision instability*.[1] In such cases, the relative merits of an agent's options vis-à-vis their causal expected utility fluctuate according to the agent's degrees of belief concerning her choices. Thus, an act that maximizes causal expected utility conditional upon a certain choice may fail to maximize this quantity conditional upon an alternative choice. When this happens, CDT is liable to offer highly counterintuitive rational (im)permissibility verdicts.

The defenders of CDT have typically sought to defuse the concerns raised by decision instability by invoking dynamical aspects of rational deliberation.[2] According to this response, CDT's potential failure to offer an agent normatively adequate recommendations at the *start* of her deliberation need not preclude the theory from offering her correct advice by the process's *end*, which is the only point at which the theory's conclusions may be reasonably put into action. While certainly an improvement over

[1] Objections of this kind were originally brought against CDT by Richter 1984, but famously renewed by Egan 2007. Recent installments in this line of attack upon CDT include Ahmed 2014a, Hare and Hedden 2016, Spencer and Wells 2019, and Spencer 2021.

[2] See, e.g., Arntzenius 2008, Joyce 2012, Joyce 2018, and Lauro and Huttegger 2020, all of whom rely upon the formal models of deliberation developed by Skyrms 1990. A similar, though distinct, defense of CDT is suggested by Bales 2020.

the unvarnished theory, this sort of deliberational CDT still fails to evade many of the troubles caused by cases of decision instability. In fact, in several of the most disturbing cases for CDT, deliberation offers little help at all. If this reply is the best causalists can muster, their theory is in a hole.

This essay tentatively ventures an alternative response on the part of the causalist that relies upon taking seriously the (in)famous thesis that *deliberation crowds out prediction*. Proponents of this thesis maintain that deliberation concerning whether to perform a certain act is incompatible with adopting any particular degree of belief in the proposition that one will in fact perform the given act.[3] One way to construe this proposal is as a norm constraining deliberating agents to adopt maximally *imprecise* or *indeterminate* credences concerning their own (immediately exercisable) options. Let us call agents who satisfy this norm *open-minded*. While existing motivations for open-mindedness are controversial, this essay can be taken as advancing a novel, pragmatic one: open-minded agents, unlike their opinionated counterparts, can reap the genuine benefits of CDT while dodging its alleged costs in cases of decision instability. Those who embrace the deliberation crowds out prediction thesis can thus maintain that the apparent poor recommendations of CDT in cases of decision instability are the fault not of CDT itself but of agents' failures to satisfy a norm of open-mindedness.

To make this case, I first introduce CDT and its attractive motivations (§2) before considering a bevy of decision problems involving instability that make trouble for it (§3). I then briefly rehearse the standard causalist response to this trouble in terms of deliberational dynamics and review its shortcomings (§4), leaving CDT in the lurch. Next, I lay the foundation for advancing a new response by introducing the deliberation crowds out prediction thesis (§5), which I subsequently develop into a full reply (§6). How far this reply can ultimately carry CDT in terms of deflecting objections stemming from decision instability will depend upon the choice rule we accept as appropriate for use by agents with imprecise credences. I will consider three such rules that each present themselves as somewhat attractive in this context: a *maximin* rule, a *permissive* rule, and a *hierarchical expected utility* rule. Ultimately, I will argue that only the third seems to yield all the intuitive rational permissibility verdicts we are after. Whichever rule we opt for, however, we can conclude that coupling CDT with open-mindedness constitutes an improvement over more familiar causalist strategies for handling decision instability, as well as over various recent alternatives to CDT designed to avoid its alleged deficiencies (§7-8).

2 Newcomb's Paradox

CDT first gained traction among philosophers as a promising replacement for its intellectual predecessor, Evidential Decision Theory (EDT), which had been freshly vitiated by *Newcomb's paradox*.[4] According to EDT, rational agents choose acts from amongst

[3] This thesis finds articulation and defense in, for example, Spohn 1977 and Levi 1993.

[4] EDT is the fruit of Jeffrey 1965/1983, while the Newcomb challenge originates in Nozick 1969. Of course, not everyone agrees that EDT was vitiated by Newcomb; for a vigorous defense of EDT and its handling of the paradox, see Ahmed 2014b.

their available options so as to maximize a *desirability function*, V, with the property that:

> $V(A) = \sum_i P(O_i|A)V(O_i)$, for any available act A, where $\{O_i\}_i$ is a partition of outcome-propositions.[5]

With P understood to be the deliberating agent's probabilistic credence function, V can be recognized as a measure of the *auspiciousness* or *news value* of an agent's potential actions. The value of any act is determined by a weighted average of the values of all possible outcomes, where each outcome's weight is fixed by how likely the agent takes it to be, conditional upon the performance of the act in question.

While a reasonable enough choice rule for many contexts, decision theorists began searching for alternatives to EDT after Newcomb's paradox revealed that the rule of V-maximization is insufficiently attentive to the causal structure of choice problems, occassionally resulting in misguided advice to seek good news in place of bringing about good results.

> **Newcomb:** A superintelligence places two boxes before you: one opaque and one transparent. The transparent contains a thousand dollars, while the opaque, you are informed, contains either one million dollars or nothing. You are offered a choice between taking both boxes or taking just the opaque box. The problem would be trivial were it not for the fact that the superintelligence has attempted to predict your choice and has (at some point in the past) placed a million dollars in the opaque box if and only if they predicted you would leave the transparent box behind.

Let us assume that you take the superintelligence to be a highly reliable (though perhaps still imperfect) predictor of your behavior and thus take your choice to be highly indicative of the prediction made. We shall also suppose that, for purposes of this problem, you only care about money and value it linearly. One formal instantiation of **Newcomb** is then given in the desirability and probability matrices represented in Tables 1 and 2.

	PredictOne	*PredictTwo*
TakeOne	1,000,000	0
TakeTwo	1,001,000	1,000

Table 1: Newcomb Desirabilities

	PredictOne	*PredictTwo*
TakeOne	0.45	0.05
TakeTwo	0.05	0.45

Table 2: Newcomb Probabilities

The information in these tables is all one needs to apply EDT and observe that taking one box is the V-maximal act in **Newcomb**.[6] But, of course, taking one box cannot possibly cause you to receive a better outcome than taking both boxes (assuming

[5] An *outcome-proposition* is a proposition strong enough to fix everything an agent cares about, i.e., a proposition such that the agent is indifferent between any two propositions that each entail it. For simplicity, I assume the finitude of outcome partitions throughout.

[6] To see this, note that $V(TakeOne) = \sum_i P(O_i|TakeOne)V(O_i) = (.9)(1,000,000) + (.1)(0) = 900,000$, while $V(TakeTwo)\sum_i P(O_i|TakeTwo)x = (.9)(1,000) + (.1)(1,000,100) = 100,910$.

you lack causal control over the past). Indeed, the choice is guaranteed to leave you a thousand dollars poorer than the alternative of taking both boxes. While **Newcomb**, as I have presented it, is admittedly fantastical, there are arguably more realistic decision problems with a similar structure.[7] If we want a decision theory that offers plausible verdicts in such cases, we need one that is sensitive to the distinction between causation and correlation in a way that EDT is not.

Enter CDT as a replacement for EDT intended to preserve its various virtues while correcting its apparent shortcomings in cases like **Newcomb**. Rather than maximizing V, CDT proposes that rational agents maximize a *utility function*, U, with the property that:

> $U(A) = \sum_j P(K_j) \sum_i P(O_i|AK_j)V(O_i)$, for any available act A, where $\{O_i\}_i$ is a partition of outcome-propositions and $\{K_j\}_j$ is a partition of dependency hypotheses.[8],[9]

The *dependency hypotheses* over which the outer sum ranges may be thought of as complete specifications of how exactly an agent's available acts causally bear upon outcomes of interest to her. In other words, a dependency hypothesis specifies the objective chance with which each of an agent's acts will result in various outcomes. For example, there are two dependency hypotheses entertained as possible by an agent facing **Newcomb**: (i) taking one box will earn you a million dollars while taking both will earn an additional thousand, and (ii) taking one box will earn you nothing while taking both will earn you a thousand. These, of course, correspond, respectively, to the state of the predictor having predicted you'd take one box and the state of the predictor having predicted you'd take both boxes (i.e., to the columns of Tables 1 and 2).

The first insight of CDT consists in recognizing that, once a dependency hypothesis is considered fixed, an agent's options each have an objectively determinate propensity to realize outcomes of various value. Assuming that the subjective probabilities of rational agents defer to chances in line with the Principal Principle, for fixed dependency hypothesis K_j and act A, this objective expected value is given by the above inner sum: $\sum_i P(O_i|AK_j)V(O_i)$.[10] In the special case where an agent is certain that K_j is the true dependency hypothesis, she ought to simply select her act so as to maximize this quantity. CDT generalizes from this starting point and suggests that when an agent is uncertain of which dependency hypothesis is true (as is typical), she should select an act so as to maximize her unconditional subjective expectation of objective expected value. This is just how U was characterized above.

To compute $U(TakeOne)$ and $U(TakeTwo)$ then, Tables 1 and 2 again provide

[7] For example, Nozick 1969 and Lewis 1979 suggest that suitable Prisoner's Dilemma scenarios may qualify. Similarly, many of the fantastical problems discussed below also enjoy well-known game-theoretic analogues. For more on the connection between exotic decision problems and familiar games, see the discussion in Weatherson MS.

[8] As in the case of outcomes, I assume agents contemplate only a finite number of dependency hypotheses throughout.

[9] There are several prominent formulations of CDT in the literature, including those of Gibbard and Harper 1978, Skyrms 1980, and Joyce 1999. The version I employ here is that of Lewis 1981.

[10] For more on the Principal Principle, see Lewis 1980.

us with all the information we need. However, in **Newcomb**, knowledge of the dependency hypotheses' precise probability values is unnecessary for concluding that $U(TakeTwo) > U(TakeOne)$ and thus that CDT correctly recommends taking both boxes. This is thanks to the fact that U-maximization, unlike V-maximization, will always respect:

> **Causal Dominance:** When selecting from a finite choice set, agents ought never select an act A if there exists another available act B such that the objective expected value of B exceeds the objective expected value of A under *every* dependency hypothesis deemed possible by the agent.

CDT respects this principle because if an act B has greater objective expected value than A under every possible dependency hypothesis, then $U(B)$ must likewise be greater than $U(A)$, regardless of how one assigns probabilities to the dependency hypotheses.[11]

Unlike EDT then, CDT appears to have the correct fundamental orientation in its approach to decision making: rational choice is about causally promoting good outcomes, not evidentially signifying them. This difference is neatly revealed in the two theories' conflicting analyses of **Newcomb**. If cases like this were the only ones that separated CDT and EDT, we could thus perhaps end our story here in favor of CDT. But, alas, **Newcomb**-like problems are not the only ones where the causal and evidential import of an agent's acts diverge.

3 Decision Instability

CDT correctly regards the evidential bearing of your available choices upon the predictions of the superintelligence as irrelevant with respect to determining which choice you should make in **Newcomb**. But discarding the evidential significance of your choices in this way, as U-maximization seems to prescribe, does not generally lead to such agreeable results. Sometimes the evidential bearing of your potential choices with respect to the antecedent causal structure of the world *does* seem like an entirely prudent factor to take into account when deliberating about what to do. Such cases typically involve the phenomenon of *decision instability*, with the optimal act to perform varying conditional upon which act you think you will actually perform. To give a sense of the problems cases of decision instability can raise for CDT, it will be helpful to introduce here three prominent cases of decision instability from the literature.

3.1 Psycho-Button

Much of philosophers' recent anxiety regarding CDT stems from the counterintuitive nature of its verdicts in **Psycho-Button**, a problem introduced by Egan 2007 but pre-

[11] In **Newcomb**, given the deterministic nature of the relevant dependency hypotheses, it is especially easy to see that CDT satisfies Causal Dominance: $U(TakeOne) = \sum_j P(K_j) \sum_i P(O_i|K_j.TakeOne)V(O_i) = P(PredictOne)(1,000,000) + P(PredictTwo)(0) = P(PredictOne)(1,000,000)$, while $U(TakeTwo) = \sum_j P(K_j) \sum_i P(O_i|K_j.TakeTwo)V(O_i) = P(PredictOne)(1,000,100) + P(PredictTwo)(1,000)$. Given that $P(PredictOne)$ and $P(PredictTwo)$ are both non-negative and at least one is positive, we may conclude that $U(TakeTwo) > U(TakeOne)$.

sented in sanitized form here:[12]

> **Psycho-Button:** A superintelligence places a button before you. The button is either rigged so that pushing it will credit a hundred dollars into your bank account or so that pushing it will debit two hundred dollars from your bank account. You are offered a choice between pushing or not pushing the button. Finally, you know that the superintelligence has (at some point in the past) rigged the button to credit the money into your account if and only if they predicted you would *not* push the button.

Let us assume that you take the superintelligence to be a highly reliable (though perhaps still imperfect) predictor of your behavior and thus take your choice to be highly indicative of the prediction made. We shall also suppose that, for purposes of this problem, you only care about money and value it linearly. One formal instantiation of **Psycho-Button** is then given in the desirability and probability matrices represented in Tables 3 and 4.

	PredictPush	*PredictDon't*
Push	-200	100
Don't	0	0

Table 3: Psycho-Button Desirabilities

	PredictPush	*PredictDon't*
Push	0.09	0.01
Don't	0.09	0.81

Table 4: Psycho-Button Probabilities

Recognizing that the superintelligence's predictions again constitute the relevant dependency hypotheses, it easy to calculate both that EDT recommends against pressing the button (i.e., $V(Don't) > V(Push)$)[13] and that CDT recommends the opposite (i.e., $U(Push) > U(Don't)$).[14] **Psycho-Button** thus seems like an advantage for EDT and a liability for CDT. Conditional upon pressing the button, one will (assuming updating by conditionalization) become very confident that the superintelligence has predicted this and thus that pressing is very likely to cause one to lose two hundred dollars. In the common jargon, pushing the button is *unratifiable*: it appears suboptimal conditional upon its own enactment. The CDT agent is thus bound to regret her decision in this case.

Of course, the same is true for agents who decide not to press the button. Conditional upon this decision, one will also become very confident that the superintelligence has predicted this and thus that pressing would very likely have caused one to gain a hundred dollars. So neither option in this case is ratifiable. Nonetheless, there still seems to be a prudentially relevant asymmetry between the two choices. One will intuitively regret pressing the button much more than one will regret abstaining, which may suggest that abstaining is the rationally obligatory course of action. Even if this common intuition is misguided, however, it seems highly plausible that we should at

[12] Egan's original presentation involves deciding whether to press a button that will kill all the psychopaths in the world, hence the case's name. This classic framing of the case, to my mind at least, invites obvious moral objections that distract from the relevant dialectic, hence my preference for discussing a sanitized version.

[13] $V(Don't) = 0 > -170 = (-200)(0.9) + (100)(0.1) = V(Push)$.

[14] $U(Push) = (-200)(0.18) + (100)(0.82) = 46 > 0 = U(Don't)$.

least avoid insisting that the opposite course of action (i.e., pushing the button) is itself obligatory and, thus, that refraining from pushing the button is not even rationally permissible. But a straightforward application of CDT to the above version of **Psycho-Button** appears to yield just this unhappy result. Hence, the case certainly seems to constitute a *prima facie* strike against CDT.[15]

3.2 Dicing with Death

An even stronger objection to CDT stemming from decision instability comes in the form of the **Dicing with Death** example cooked up by Ahmed 2014a:[16]

> **Dicing with Death:** A superintelligence places two buttons before you, *A* and *B*. The buttons are rigged so that exactly one of them will debit a thousand dollars from your bank upon being pressed and exactly one will have no effect on your finances upon being pressed. You are offered a choice between pressing *A* or pressing *B* or paying a dollar to flip a fair coin that will trigger *A* if it lands heads and trigger *B* if it lands tails. Finally, you know that the superintelligence has (at some point in the past) rigged button *A* to debit the money from your account if they predicted you would press *A* and has rigged button *B* to debit the money from your account if they predicted you would press *B*, but, possessing no ability to predict the outcome of your coin flip, has employed their own fair randomizing device to decide how to rig the buttons if they predicted you would employ yours.

Let us assume that you take the superintelligence to be a highly reliable (though perhaps still imperfect) predictor of your behavior and thus take your choice to be indicative of the prediction made in the relevant ways. We shall also suppose that, for purposes of this problem, you only care about money and value it linearly. One formal instantiation of **Dicing with Death** is then given in the desirability and probability matrices represented in Tables 5 and 6.

	RiggedA	*RiggedB*
A	-1000	0
B	0	-1000
Coin	-501	-501

Table 5: Dicing with Death Desirabilities

	RiggedA	*RiggedB*
A	0.35	0.05
B	0.05	0.35
Coin	0.1	0.1

Table 6: Dicing with Death Probabilities

Note that since **Dicing with Death** involves chances in a non-degenerate way (unlike **Newcomb** and **Psycho-Button**), the entries of Table 5 do not necessarily correspond to direct values of outcomes but rather to the objective expected value of each act-state conjunction.[17] Nonetheless, recognizing the states specifying how the buttons

[15] It is worth noting, however, that the intuitions at play in common reactions to **Psycho-Button** have been challenged by some causalists, including Joyce 2012, Armendt 2019, and Williamson 2021.

[16] I again sanitize Ahmed's presentation of the problem slightly.

[17] Of course, this remark about non-degeneracy is only relevant for the third row act, *Coin*. In the first two rows, objective expected value coincides with the value of a determinate outcome as in the previous cases.

are rigged as dependency hypotheses, we can again compute the recommendations of both EDT and CDT in **Dicing with Death**.

Doing so reveals that EDT recommends flipping the coin (since $V(Coin) > V(A) = V(B)$),[18] while CDT recommends either hitting button A or hitting button B (since $U(A) = U(B) > U(Coin)$).[19] Whatever we think of CDT's recommendation in **Psycho-Button**, its verdict here certainly seems more problematic for the theory. You are confident that if you choose either A or B you will have been predicted and suffer a loss of a thousand dollars, while you are confident that randomizing at least gives you a decent shot at avoiding this loss, yet there is no price you would pay (no matter how small) to take it. A disturbing result indeed!

3.3 Frustrator

A structurally similar counterexample to CDT has been suggested recently by Spencer and Wells 2019, who label it the **Frustrator** problem:

> **Frustrator:** A superintelligence places two boxes before you, A and B, along with an envelop E containing forty dollars. You are offered your pick of the three. The superintelligence has distributed a hundred dollars between A and B, according to their prediction of your choice. If they predicted you'd take A, the hundred is in B. If they predicted you'd take B, the hundred is in A. If they predicted you'd take the envelop, they distributed the money evenly between A and B, placing fifty dollars in each.

Let us assume that you take the superintelligence to be a highly reliable (though perhaps still imperfect) predictor of your behavior and thus take your choice to be indicative of the prediction made in the relevant ways. We shall also suppose that, for purposes of this problem, you only care about money and value it linearly. One formal instantiation of **Frustrator** is thus given in the desirability and probability matrices represented in Tables 7 and 8.

	PredictA	*PredictB*	*PredictE*
A	0	100	50
B	100	0	50
E	40	40	40

Table 7: Frustrator Desirabilities

	PredictA	*PredictB*	*PredictE*
A	0.4	0.01	0.01
B	0.01	0.4	0.01
E	0.01	0.01	0.14

Table 8: Frustrator Probabilities

Frustrator has a similar flavor to **Dicing with Death**, save that the role of the 'safe' option has been wrested from a chance device and handed over to the sure outcome of forty dollars. Similar remarks apply, however. EDT recommends taking the envelop (since $V(E) > V(A) = V(B)$),[20] while CDT recommends taking either box A or box

[18] $V(Coin) = -501 > -875 = (-1000)(0.875) + (0)(0.125) = V(A) = V(B)$.
[19] $U(A) = U(B) = (0.5)(-1000) + (0.5)(0) = -500 > -501 = U(Coin)$.
[20] $V(E) = 40 > 3.57 \approx (0)(\frac{1}{42}) + (100)(\frac{1}{42}) + (50)(\frac{1}{6}) = V(A) = V(B)$.

B (since $U(A) = U(B) > U(E)$).[21] While none of your options here are ratifiable (as in the previous two problems), taking the sure forty dollars certainly seems like the prudent thing to do, yet CDT forecloses the possibility of rationally taking this course of action. Another problem for CDT.

4 Deliberational CDT

We might distinguish two kinds of errors CDT seems to commit in the above cases of decision instability. First, CDT appears to be too restrictive an account of rational choice in the sense that it *undergenerates* rational permissibility verdicts. Refraining from pushing the button in **Psycho-Button**, for example, seems eminently reasonable and yet, given the quantities specified in Tables 3 and 4, CDT forbids this choice as irrational. Second, CDT also appears to be too loose an account of rational choice in the sense that it *overgenerates* rational permissibility verdicts. Refusing to pay a dollar to randomize in **Dicing with Death**, for example, seems patently irrational and yet CDT licenses doing so.

The most sophisticated defenders of CDT would object, in part, to the accuracy of this set of charges. Thus far, we have characterized CDT as a decision theory that moves directly from any choice problem and relevant pair of probability and desirability functions to straightforward recommendations for/against actions. However, proponents of *Deliberational CDT* charge that this is an inappropriate and short-sighted way to apply the machinery of CDT. In cases of decision instability, our deliberation typically generates information regarding which act we will ultimately choose and hence which acts are optimal to choose. A rational agent, according to Deliberational CDT, thus ought to revise her beliefs as she deliberates to take advantage of this information and put it to work in her efforts to maximize U.

Deriving from the landmark work of Skyrms 1990, there are many models that sketch what the dynamics of a deliberational process attentive to informational feedback in this way might look like. The key idea behind all of them is that an agent ought to revise her beliefs in her own act propositions in light of her (causal) expected utility calculations according to a manner that *seeks the good*. That is, upon estimating the U-values of her available options, a rational agent ought somehow to raise her degree of belief that she will take each of the options she has estimated to have an above average utility, without similarly raising her degree of belief in any of the below average utility options. The *equilibria* or fixed points of such a dynamics will be those belief states in which only U-maximal acts receive positive probability.

The probability assignment recorded in Table 4, which we employed to conclude that the utility of pushing the button outstrips that of refraining in **Psycho-Button**, is clearly not a deliberational equilibrium, since $U(Push) > U(Don't)$ yet $P(Don't) \neq 0$. Proponents of Deliberational CDT will thus insist that this is an inappropriate state of mind from which to use CDT to justify pushing the button, as it still hides unmined evidence regarding the causal import of pushing/refraining that must be factored into

[21] $U(A) = U(B) = (0)(0.42) + (100)(0.42) + (50)(0.16) = 50 > 40 = V(E)$.

your deliberation before you can use CDT to make a final, fully informed choice. In this particular example, realizing that $U(Push) > U(Don't)$ ought to lead you to adjust your degree of belief that you will push the button upward to some degree, which will in turn lessen the utility of pushing, assuming that your confidence in the superintelligence's predictive powers continues to hold up.[22] If your deliberational dynamics seeks the good and is sufficiently tempered so as to avoid wild swings in probability of the sort that could leave you oscillating between cyclic belief in pushing and in refraining, you will ultimately converge to the only stable belief state possible in this version of **Psycho-Button**, i.e., one in which your probabilities are as given in Table 9.[23]

	PredictPush	*PredictDon't*
Push	0.2628	0.0292
Don't	0.0708	0.6372

Table 9: Psycho-Button Equilibrium Probabilities

If we recompute CDT's recommendations from the standpoint of these probabilities, we find that $U(Push) = U(Don't)$.[24] Hence, the charge that CDT forbids you to refrain from pushing in **Psycho-Button** may, from this perspective, be dismissed. A careful agent who follows CDT's prescriptions only after she has factored into her calculations all information gleanable from her own deliberation is not, after all, forbidden from doing what most of us would be inclined to do in **Psycho-Button**.

However, this causalist reply does nothing to mitigate the worry that CDT overgenerates permissibility verdicts in **Psycho-Button**, since the unique deliberational equilibria that renders refraining U-maximal equally renders pushing U-maximal. Deliberational CDT lacks the resources to say that pushing is irrational in **Psycho-Button**. What's worse, Deliberational CDT is of no help at all in mitigating the force of the objections to CDT stemming from **Dicing with Death** and **Frustrator**. It is readily verifiable that the only deliberational equilibria that respect the rigidity of the conditional probabilities in these problems are as given in Tables 10 and 11.[25]

	RiggedA	*RiggedB*
A	0.4286	0.0714
B	0.0714	0.4286
Coin	0	0

Table 10: Dicing with Death Equilibrium Probabilities

	PredictA	*PredictB*	*PredictE*
A	0.4750	0.0125	0.0125
B	0.0125	0.4750	0.0125
E	0	0	0

Table 11: Frustrator Equilibrium Probabilities

From the standpoint of the epistemic states encoded by these equilibrium probabilities, it remains the case that CDT forbids the intuitively correct choices of flipping the

[22] That is, assuming that your conditional probabilities of states given acts remain *rigid* or fixed throughout the dynamics.

[23] Probabilities are approximated to the fourth decimal place.

[24] $U(Push) = (0.\overline{3})(-200) + (0.\overline{6})(100) = 0 = U(Don't)$.

[25] Figures are again rounded to the fourth decimal place.

coin in **Dicing with Death** and opting for the envelope in **Frustrator**.[26] Thus, while proponents of Deliberational CDT may be right that their proposal constitutes an improvement over a less carefully applied CDT, it is a modest improvement at best. If we want to patch up CDT so as to escape its most alarming embarrassments, we must look elsewhere.

5 Deliberation Crowds Out Prediction?

Our discussion thus far has assumed that the agents facing **Newcomb**, **Psycho-Button**, etc. have sharp credences of the sort representable as real-valued probabilities concerning every Boolean combination of act propositions and dependency hypotheses. In particular, all of the probability tables introduced so far fix probabilities for act propositions like taking one box, pushing the button, etc. A prominent line of decision theorists, most notably Spohn 1977 and Levi 1993, have strenuously objected to this assumption.[27] According to these theorists, *deliberation crowds out prediction*.[28] By this, these authors mean to assert that a deliberating agent's potential choices, as matters under her immediate control, are inappropriate objects to subsume under the domain of her credence function. On this view, while conditional degrees of belief *given* one's acts are rationally legitimate attitudes for a deliberating agent to adopt, unconditional degrees of belief in her acts themselves are ruled out.[29]

There are various motivations that have been offered in favor of this point of view. For starters, the idea of ascribing act credences to deliberating agents is incongruent with the traditional operationalization of credence in terms of betting behavior. If I set my fair betting quotient on an act proposition to any non-extreme number between 0 and 1, I am setting myself up for needless loss, while if I already set it to either of the extreme values of 0 or 1 it is perhaps unclear in what sense I can be said to be truly deliberating.[30] As seems to be generally recognized today, however, this argument is far from decisive since the brute identification of credences with fair betting rates is commonly recognized as too crude, particularly in cases of moral hazard in which the activity of betting is itself liable to influence the likelihood of the propositions being bet on.[31]

Another concern, famously emphasized by Spohn, is that act credences are pragmatically pointless, allegedly playing no significant role in the modelling of a rational

[26] This general conclusion is granted and defended by Deliberational CDT's foremost advocate in Joyce 2018.

[27] See also Gilboa 1999, Price 2007, Levi 2000, and Levi 2007.

[28] More recently, Hajek 2016, in a critical discussion, has dubbed this thesis *Deliberation Annihilates Reflexive Credences (DARC)*.

[29] Spohn 2012, Spohn (unpublished) develop a subtler and more sophisticated variant of the thesis that involves drawing a sharper distinction between *action* and *decision* variables than has commonly been drawn in the decision theory literature. Unfortunately, I lack space to adequately deal with Spohn's more recent proposals here.

[30] See Spohn 1977 and Levi 2007 for variations on this argument.

[31] For prominent critiques of the betting argument against act credences, see Rabinowicz 2002 and Hajek 2016.

agent's deliberative processes. This is arguably correct in the case of EDT, whose decision rule instructs agents to consult their conditional probabilities given acts but never unconditional act probabilities in the course of assessing the relative instrumental value of available acts. As long as we resist the reduction of conditional to unconditional probability via the Ratio Rule, EDT thus indeed carries with it no commitment to act probabilities.[32] However, the charge of pragmatic epiphenomenalism carries less water when directed against agents that subscribe to the version of CDT we have been considering. While CDT's decision rule does not directly invoke act probabilities, it does employ unconditional probabilities in dependency hypotheses. When conjoined with conditional probabilities for such hypotheses given acts, these may suffice to fix determinate act probabilities since for any dependency hypothesis K and act A, the law of total probability requires that $P(K) = P(K|A)P(A) + P(K|\overline{A})(1 - P(A))$. Hence, given $P(K)$ (which is required for the application of CDT) and $P(K|A), P(K|\overline{A})$ (which are plausibly required for Bayesian updating), the value of $P(A)$ is also settled unless perfect act-state independence obtains (i.e., unless $P(K|A) = P(K|\overline{A})$), in which case unconditional act probabilities are indeed left open.[33]

In the cases that form our present concern (i.e., those involving probabilistic act-state dependence like **Newcomb**, etc.), act probabilities thus need not be epiphenomenal in the case of CDT agents. Act probabilities can, it turns out, indirectly influence a CDT agent's preferences via their logical connection to state probabilities of the sort that causalists maintain are directly relevant for computing causal efficacy values. (E.g., if you think you are likely to press the button in **Psycho-Button**, then you must also think you are likely in the state of the world where pressing will cause bad outcomes and hence where you shouldn't press the button, etc.[34]) However, this evasion of epiphenomenalism is at best a Pyrrhic victory for the defender of act credences. Such credences may indeed influence behavior, but generally they do so to the deliberating agent's *detriment*. Or so I wish to argue. The apparent errors of CDT in the puzzle cases rehearsed above can be traced to the adoption of determinate act credences. If I am right, the problem with self-prediction is then not so much that it is either useless or senseless, but more simply and disturbingly that it seems to impair rather than aid rational deliberation.[35],[36]

[32] Of course, the classic formulation of EDT in Jeffrey 1965/1983 does involve act probabilities, but this is not essential to EDT as such.

[33] It was, of course, problems like this that decision theories like that of Savage 1954/1972 attempt to model and that Spohn no doubt had in mind when making his argument.

[34] This pragmatic relevance of sharp act credences for CDT is noted by Hajek 2016 and Podgorski 2022, the latter of whom notes that this fact puts CDT at odds with a principle previously endorsed by Joyce 2002, according to which "...it is absurd for an agent's views about the advisability of performing any act to depend on how likely she takes that act to be." (Joyce 2002, p. 79)

[35] This essay may thus be seen as taking up a challenge to Spohn's view posed by Rabinowicz 2002: "Even if it were true that as deliberators we have no use for the probabilities of the options among which we choose, Spohn would still need to show that such probabilities would be positively harmful." (Rabinowicz 2002, p. 113)

[36] Joyce 2002 has offered an argument in defense of act credences against the charge of epiphenomenalism even in cases where they play no evaluative role in an agent's assessment of the (expected) utility of acts on the grounds that act credences play a crucial role in causally explaining rational agents' behavior: a free agent must take her belief that she will perform a given act to be causally efficacious in the sense of causing her to perform it. This argument, if sound, seems to me to support act credences only following the conclusion

6 Imprecise CDT

One way to precisfy the deliberation crowds out prediction thesis is along the lines of a norm of *open-mindedness* stated in terms of *imprecise* or *indeterminate* probabilities:

> **Open-mindedness:** A rational agent deliberating with respect to a set of available acts $\mathscr{A}$ ought to have maximally indeterminate credences with respect to the members of $\mathscr{A}$, i.e., her doxastic state ought to be representable by a set $\mathscr{P}$ of probability measures whose individual marginalizations with respect to $\mathscr{A}$ collectively yield the set of all possible probability distributions over $\mathscr{A}$.

An agent whose degrees of belief are representable by a set $\mathscr{P}$ of probability measures (known as a *credal representor* comprised of *avatars*) is epistemically committed to exactly those likelihood judgments shared in common by $\mathscr{P}$'s members.[37] For example, such an agent ascribes a sharp probability of x to a proposition p just in case all members of $\mathscr{P}$ (i.e. all her avatars) assign p probability x. Similarly, she assigns a probability within a range R to proposition p just in case each of her avatars assigns x a probability within R. I will refer to the intersection of all ranges R such that an agent with credal representor $\mathscr{P}$ assigns p a probability within R as *the* credence an agent with $\mathscr{P}$ assigns to p. In the examples of interest to us, credences will generally either be points (sharp probabilities) or non-trivial intervals (imprecise probabilites). Relative to traditional probability theory, this imprecise framework has the advantage of enabling us to model reasonable indeterminacy in an agent's degrees of belief, i.e., contexts where an agent can't sensibly judge either the absolute or relative likelihood of propositions. According to **Open-mindedness**, deliberation provides one such context: a deliberating, open-minded agent makes no judgments concerning how likely she is to choose any of the options currently falling under the purview of her deliberation.[38]

Note again that the causalists brought to ruin above by **Psycho-Button**, **Dicing with Death**, and **Frustrator** all violated **Open-mindedness** by adopting sharp act-credences. Can we reasonably argue they would have done better had they instead abided by this norm? To answer this question, we of course first need an account of how CDT ought to be applied in contexts where agent's lack sharp probabilities, since the formulation of CDT offered above assumes credal precision. I find at least three proposals in this regard worth considering: (i) *Maximin CDT*, (ii) *Permissive CDT*, and (iii) *Hierarchical CDT*. The first and second of these decision rules are attractive in various ways but ultimately unacceptable in my view (though readers may disagree!). The third is, I believe, substantially more promising.

---

of deliberation, when the time for decision has come, rather than during the midst of its process, but I lack space to give an adequate treatment of Joyce's important work here.

[37] For a pair of nice introductions to imprecise probabilities, see Mahtani 2019 and Bradley 2019.

[38] Stern 2018 also briefly considers an interpretation of DCOP in terms of imprecise probabilities, though not in the exact form presented here. While the key proponents of DCOP (e.g., Spohn and Levi) never themselves cast their thesis in exactly this way, **Open-mindedness** seems an entirely reasonable way to capture the spirit of their view. To hold no doxastic commitments at all towards a proposition (beyond those required by formal consistency) is simply to be representable as having maximally imprecise credences with respect to it. To hold no doxastic commitments with respect to one's act propositions, as required by DCOP, is then simply to satisfy **Open-mindedness**.

6.1 Maximin CDT

Suppose that you agree with the guiding thought behind CDT that instrumental value is a matter of causal efficacy rather than auspiciousness, but that you also satisfy **Open-mindness** and thus adopt maximally indeterminate act credences. For concreteness, let's say that the decision problem you currently face is a version of **Psycho-Button** with sharp *conditional* probabilities as given in Table 4. Each avatar in your representor can apply traditional CDT to arrive at causal efficacy values for pushing and not pushing the button. If all of these avatars agreed concerning the relative causal merits of pushing and not pushing, we could simply say that you should respect their unanimous opinion. (This is what happens for an open-minded agent confronting **Newcomb**.) But they don't. Avatars that are relatively confident of your opting to push the button will recommend against that course, while those relatively confident of your opting not to push will recommend the opposite. An open-minded agent's assessment of the causal efficacy of her options in *Psycho-Button* is thus indeterminate. What to do?

A popular suggestion for how to go about making decisions in contexts involving this sort of credal indeterminacy is to apply the rule of *maximizing minimum expected utility*:

> **Maximin CDT:** Rational agents deliberating over a finite action set $\mathscr{A}$ ought to choose an act $a \in \mathscr{A}$ that maximizes $min_{P \in \mathscr{P}}\{U_P(a)\}$, where $\mathscr{P}$ is the deliberating agent's credal representor and U_P is causal utility relative to P.[39]

Maximin CDT clearly agrees with traditional CDT in cases where the latter applies (i.e. when $\mathscr{P}$ is a singleton), but generalizes to cover cases of credal indeterminacy as well. In such cases, Maximin CDT encodes a kind of pragmatic pessimism: if the relative causal expected utility of two acts is indeterminate, assume the worst about each and maximize the minimum. There are plenty of objections one can raise to this as a general decision rule, but, to its credit, when coupled with **Open-mindedness**, Maximin CDT does yield precisely the correct answers in each of the four cases that have concerned us so far: **Newcomb**, **Psycho-Button**, **Dicing with Death**, and **Frustrator**.

Maximin CDT gets **Newcomb** right because, even for an open-minded agent, the causal expected utility of two-boxing is determinately greater than the causal expected utility of one-boxing. Though **Psycho-Button** jettisons this determinacy, Maximin CDT still continues to yield the intuitive recommendation of not pushing the button in the case of open-minded agents since for such agents: $min_{P \in \mathscr{P}}\{U_P(Don't Push)\} = 0 > -500 = min_{P \in \mathscr{P}}\{U_P(Push)\}$. Even **Dicing with Death** and **Frustrator**, which were the greatest source of trouble for CDT in the absence of **Open-mindedness**, are handled appropriately by Maximin CDT, which unambiguously recommends, respectively, flipping the coin and taking the envelope in these problems. For example, in **Frustrator** the causal expected utility of taking the envelope for an open-minded agent is determinately 40, while, relative to distinct avatars, the causal expected utility of taking box A and taking box B can each drop to as low as 0.

[39] A rule like this is famously explored by many authors, including Gärdenfors and Sahlin 1982 and Gilboa and Schmeidler 1989.

So far, so good for Maximin CDT. Unfortunately, however, this rule's seeming virtues are largely an artifact of the puzzle cases we have so far chosen to set our focus upon. In cases like **Psych-Button**, each of the available options is self-frustrating or (in the common jargon) unratifiable. In many problems that have this feature, Maximin CDT seems to yield intuitive results, but there are other problems where its verdicts are more clearly problematic.

> **Nice Psycho-Button:** A superintelligence places a button before you. The button is either rigged so that pressing it will credit a hundred dollars into your bank account or so that pressing it will debit two hundred dollars from your bank account. You are offered a choice between pressing or not pressing the button. Finally, you know that the superintelligence has (at some point in the past) rigged the button to credit the money into your account if and only if they predicted you *would* press the button.

Let us assume that you take the superintelligence to be a highly reliable (though perhaps still imperfect) predictor of your behavior and thus take your choice to be highly indicative of the prediction made. We shall also suppose that, for purposes of this problem, you only care about money and value it linearly. One formal instantiation of **Nice Psycho-Button** is then given in the desirability and conditional probability matrices represented in Tables 12 and 13. (Note: Table 13 encodes conditional probabilities for states given acts, rather than unconditional probabilities, so as to avoid contradicting **Open-mindedness**.)

	PredictPress	*PredictDon't*
Press	100	-200
Don't	0	0

Table 12: Nice Psycho-Button Desirabilities

	PredictPress	*PredictDon't*
Press	0.9	0.1
Don't	0.1	0.9

Table 13: Nice Psycho-Button Conditional Probabilities

For the open-minded agent, **Nice Psycho-Button** again presents a problem involving indeterminate causal efficacy judgments. However, in this example, your available choices are self-fulfilling rather than self-frustrating. For example, conditional upon pushing the button, you are likely to walk away with a hundred dollars more than you would have gotten by not pushing it. It now seems intuitive that pushing the button is rationally permissible and perhaps even obligatory. Yet Maximin CDT treats this case exactly the same as the original **Psycho-Button**: it is still refraining that uniquely maximizes minimum causal expected utility. We thus have a relatively clear case of Maximin CDT under-(and likely over-)generating rational permissibility verdicts. This seems enough to set aside Maximin CDT as a generally applicable account of rational choice.

6.2 Permissive CDT

A more popular and seemingly well-motivated approach to decision making using imprecise probabilities involves taking a decidedly more liberal attitude than that suggested by Maximin CDT. Before stating this second rule, let's first define that an act a has a causal efficacy value determinately greater than that of act b, written $a \succ b$, for an agent with credal representor $\mathscr{P}$, just in case the causal expected utility of a is greater than that of b relative to each $P \in \mathscr{P}$. An agent's $\succ$-maximal options are those that are maximal with respect to the partial ordering on acts induced by $\succ$. We can then propose:

> **Permissive CDT:** Rational agents deliberating over a finite action set $\mathscr{A}$ ought to choose any $\succ$-maximal act in $\mathscr{A}$.

Like Maximin CDT, Permissive CDT clearly collapses to traditional CDT when all relevant probabilities are sharp. However, unlike Maximin CDT, Permissive CDT never seems to under-generate permissibility verdicts when **Open-mindedness** is satisfied. In each of the examples considered so far, all of the intuitive choices are deemed rationally appropriate by Permissive CDT applied to open-minded agents (e.g. two-boxing in **Newcomb**, refraining in **Psycho-Button**, pushing in **Nice Psycho-Button**, etc.). From a thorough-going causalist perspective, this can be expected to hold true in general. Permissive CDT only rules out acts that are determinately inferior to alternatives with respect to causal expected utility. But, from a causalist standpoint, any act that is determinately inferior to another in this way can never be choiceworthy.

Unfortunately, the liberalism underwriting this virtue of Permissive CDT also intuitively leads to significant over-generation of permissibility verdicts. For example, pushing the button in **Psycho-Button** and refusing to randomize in **Dicing with Death** are not ruled irrational by Permissive CDT in the case of open-minded agents. Given the desiderata I have set out for ourselves, this is a heavy strike against the theory. Still, the strike may be less decisive than the case against Maximin CDT. All of the options I (and likely you) would select in **(Nice) Psycho-Button**, **Dicing with Death**, etc. are considered perfectly acceptable by Permissive CDT (assuming we are open-minded), but not by Maximin CDT. So, in rationally reaching my decisions, I am never forced to contradict Permissive CDT's recommendations as I am Maximin CDT's. Moreover, in many cases where I choose intuitively (e.g. by refraining in **Psycho-Button** or taking the envelope in **Frustrator**), I will often then judge it to be the case that had I made an alternative decision also licensed by Permissive CDT (e.g. pushing in **Psycho-Button** or taking a box in **Frustrator**), I would have actually been better off. So perhaps I should not hold it against Permissive CDT that it would have licensed me to accrue those rewards! No doubt these alternative decisions still seem ex ante irrational to many of us, but perhaps the sort of credal indeterminacy regarding causal structure at play for open-minded agents in these examples should actually make these alternatives seem somewhat rationalizable.

Be that as it may, with Permissive CDT as our rule for decision making in contexts of credal indeterminacy, I believe we can already recognize **Open-mindedness** as an attractive constraint upon rational deliberation, affording us all the rational license we

could reasonably want. However, I concede that I still find it troubling that this combination also yields, alongside this freedom, a good number of additional permissibility verdicts we (at least intuitively) don't want. If we decide our impermissibility intuitions are as firm as our permissibility ones in these cases after all, is there a simple decision rule that might allow **Open-mindedness** to more fully save our intuitions regarding these cases?

6.3 Hierarchical CDT

There is. Or so it seems to me, though I should concede up front that I am not entirely confident of how theoretically well-grounded the proposal ultimately is. Its main virtue is simply that it does, by my lights at least, yield intuitively correct (im)permissibility verdicts across the range of decision problems involving act-state dependence. To achieve this result, the rule I propose supplements the verdicts of Permissive CDT with those of standard EDT in a hierarchical fashion:

> **Hierarchical CDT:** Rational agents deliberating over a finite action set $\mathscr{A}$ ought to choose any V-maximal member of the set of $\succ$-maximal acts in $\mathscr{A}$.

Again, in the presence of sharp act credences, Hierarchical CDT is nothing other than standard CDT. But, in the case of agents with imprecise credences (e.g., open-minded agents), Hierarchical CDT effectively instructs agents to first apply Permissive CDT and rule out all determinately causally inferior options (e.g., taking one box in **Newcomb**) and then amongst the surviving options choose one that maximizes desirability.

This theory handles all the problems so far considered (again assuming **Open-mindedness**). It uniquely yields refraining from pushing the button in **Psycho-Button**, pushing the button in **Nice Psycho-Button**, randomizing in **Dicing with Death**, and taking the envelope in **Frustrator**. And it can, I think, be given a sensible enough rationale, by causalist lights, even though I lack the sort of complete justificatory story that one could feel confident in proposing as a fully satisfying theoretical grounding. From the causalist standpoint, rational choice aims at bringing about the best outcomes one can given the causal structure of the world and one's uncertainty regarding it. That is, a causalist agent wants to perform acts of greatest possible causal efficacy with respect to realizing her ends, i.e., acts that maximize U. It thus stands to reason that when the causal efficacy value of one act is determinately inferior to another (e.g., as in **Newcomb**), the determinately lesser act may safely be removed from deliberative consideration. But when an agent's credal state fails to fix any determinately U-optimal acts, this elimination of causally inferior acts is insufficient to draw deliberation to its close. How should an agent choose amongst acts whose relative causal efficacy values are indeterminate? It seems to me that turning to considerations of auspiciousness à la EDT, while inappropriate as an account of the fundamental aim or orientation of rational choice, is nonetheless at the very least as reasonable a way as any to make decisions when more fundamental causal considerations fall silent.

But we can say more. There is perhaps an analogy to be drawn here between the role played by evidentialist considerations in Hierarchical CDT and the role played

by intentions in fixing the rationality of behavior according to some action theorists, most notably Bratman 2012.[40] As Bratman and others rightly emphasize, intentions lack the central action-licensing significance of more fundamental decision guiding attitudes like degrees of belief and value judgments. The mere fact that one intends to ϕ does nothing to render ϕing rational if ϕing is not independently recognizable as an instrumentally effective course of action. However, it would be wrong to infer from this that intentions have no role to play in setting bounds for rational action. Supposing that multiple available courses of action are judged rationally adequate relative to one's beliefs and aims, prior intentions to act in certain ways can play a critical role in arbitrating amongst competing practical possibilities. Bratman has convincingly argued that ascribing intention such a role in the theory of rational action allows agents to coordinate their behavior, both intertemporally and interpersonally, in ways that would otherwise be closed to them.[41] If this view is correct, instrumentally rational action has a hierarchical structure to it: most fundamentally, rationality in action is determined by the acting agent's degrees of belief and value judgments (i.e., by considerations of expected utility), but, at a secondary-level, such rationality is also constrained by considerations of intention and planning.

Analogously, the proponent of Hierarchical CDT has a similarly layered view of rational choice. Fundamentally, it is determined by causal considerations, but evidential considerations may still play a secondary role in rational arbitration. There may, in fact, be more than just a faint analogy here between Hierarchical CDT and the view that intention can act as a filter on rationally admissible choices. In the decision problems considered hereto, agents have not been afforded the opportunity to form intentions to pre-commit to particular choices causally prior to the predictions of the superintelligence. But we can nonetheless ask ourselves what intentions they would have been rational (in a causalist sense) to form in this regard if they could have done so. Invariably in these examples, the recommendation of CDT would be to pre-commit to acting in accord with EDT. For example, in **Newcomb**, one wants to lead the superintelligence to place a million dollars in the opaque box, so pre-committing to one-boxing is advisable on causalist grounds. In **Psycho-Button**, one knows that pre-committing to pushing the button will lead the superintelligence to configure the button so as to debit two hundred dollars from your account, and so the only sensible thing to do is to pre-commit to refraining. And so on. Thus, in these particular cases, the prescription to maximize desirability among causally admissible options is tantamount to deferring to suitable counterfactual pre-commitment intentions when the verdicts of causal expected utility maximization are ambiguous.[42]

[40] See also Bratman 1987. [Include also remark redacted for anonymity.]

[41] The simplest examples of what I take to be Bratman's point involve value incommensurability. Suppose I judge α and β to be incomparable in value. Then there may exist an alternative outcome γ such that I determinately prefer α to γ, though I still judge β to be incomparable to γ. (Perhaps, α is a career as a doctor, β a career as a lawyer, and γ a career as a doctor less five dollars.) If I must decide between α and a choice between β and γ, it seems I have only two possible rational courses of action: opting for α or rejecting α and then selecting β. But this implies that there is something irrational about selecting γ at the second stage of the problem, a verdict most naturally accounted for in terms of the irrationality of antecedently intending to pick γ.

[42] Spohn 2012 has gone further and suggested that agents facing genuine Newcomb problems may be thought of as in fact having implicit prior intentions to act in certain ways, intentions that are then born out

As indicated, I don't pretend that a story like this constitutes a fully satisfying theoretical foundation for Hierarchical CDT. Too many questions remain. For example, even granting that actual intentions can act as rational filters in contexts of evaluative indeterminancy, why accord the same status to counterfactual ones? Moreover, while it seems intuitive that EDT coincides with appropriate causalist pre-commitment dispositions in the sorts of cases that concern us, I haven't worked out this suggestion in any precise detail, let alone offered any general proof of the hypothetical identification. Still, I think we have said enough to warrant taking Hierarchical CDT seriously as a candidate for open-minded application of CDT. Additionally, as we shall see below, adopting this method of resolving evaluative indeterminacy allows us to satisfy certain plausible rationality principles that other approaches (e.g. Deliberational CDT) have notably foundered upon. My central argument on the approach's behalf, however, remains the fact that, when coupled with **Open-mindedness**, Hierarchical CDT saves the relevant phenomena, so to speak, by neatly systematizing a wide range of our intuitive rational permissibility judgments.

The import of all of this for our defense of **Open-mindedness** should be clear: if Hierarchical CDT is the right way for agents to make decisions in the face of credal indeterminacy, then a strong pragmatic argument can indeed be made in favor of **Open-mindedness**. Open-minded agents following Hierarchical CDT never seem to make worse decisions than their more opinionated, sharp-credenced counterparts, and sometimes do much better. Or rather, this is close to the truth. Hierarchical CDT still requires, in my view, a modest revision before I can fully stand behind my claim that it constitutes the correct account of rational choice. But this revision is most easily introduced in the context of comparing it with other exisiting accounts of rational choice, and so I shall defer it until §7.3.

7 Alternative Theories?

The strength of this argument for **Open-mindedness** of course depends somewhat on the availability of other means to reach similarly agreeable results. I will thus conclude my argument for **Open-mindedness** by canvassing some other recent attempts to correct CDT's apparent flaws in handling problems involving decision instability and compare their merits to those of endorsing **Open-mindedness**. I shall assume throughout that Hierarchical CDT offers the correct account of decision making in contexts of credal indeterminacy, though my comments should leave it fairly obvious how Maximin and Permissive CDT compare to the surveyed theories as well.

7.1 Deliberational CDT Revisited

The advantage of coupling CDT with the **Open-mindedness** postulate relative to supplementing it with deliberational considerations of the sort suggested by authors like Skyrms, Arntzenius, and Joyce, should by now be apparent. The former can allow us to account for the rationality of such actions as randomizing in **Dicing with Death**

in their behavior in such problems.

and taking the envelope in **Frustrator**, while the latter cannot. At the same time, embracing **Open-mindedness** is compatible with preserving the central virtues of Deliberational CDT, most notably its satisfaction of **Causal Dominance** and its attention to the causal information gleanable from hypothetical decisions (albeit that deliberative and open-minded, hierarchical CDT agents pay attention to such causal information in quite different ways).

7.2 Graded Ratifiability

Rather than drawing radical epistemic lessons like **Open-mindedness**, some philosophers have taken the cluster of decision problems discussed so far to motivate constructing entirely new decision rules for assessing the choiceworthiness of acts. Perhaps most famously, Harper 1986, building on an idea of Jeffrey 1965/1983, suggests supplementing standard CDT with a ratifiability condition: maximize U only amongst ratifiable options, i.e., options that maximize U conditional on their own performance. The obvious objection to such a proposal is that some decision problems (including **Psycho-Button**, **Dicing with Death**, etc.) lack ratifiable options. Harper replies that such problems only lack ratifiable options if we rule out the possibility of *mixed strategies*, that is, random choices employing (internal or external) chance devices. When such genuine randomization is possible and unpenalized, this move suffices to render Harper's proposal applicable. But if these conditions can't be met (e.g., if randomizing is either impossible or costs something, as in **Dicing with Death**), ratifiable options may again disappear, rendering the ratifiability criterion of little use.[43]

In light of the potential inapplicability of the absolute ratifiability criterion, Barnett (forthcoming) draws our attention to the notion of *graded ratifiability*. Even in decision problems where no strictly ratifiable options exist, some options may still appear less ratifiable than others. For example, as noted in our initial discussion of **Psycho-Button**, while you will regret both pushing and refraining, conditional on each act's own performance, there is a sense in which you will regret pushing much more than you will regret refraining. After all, pushing is likely to cause you to lose two hundred dollars, while refraining only indicates missing out on the opportunity to receive a hundred. Formally, Barnett defines an act A's degree of ratifiability, relative to an alternative B, as $U_A(A) - U_A(B)$, where U_A is simply causal expected utility computed relative to $P(\cdot|A)$, i.e., the agent's credence function conditioned upon A. It is straightforward to verify that, so understood, the degree of ratifiability of pushing in **Psycho-Button** (relative to refraining) is indeed less than the degree of ratifiability of refraining (relative to pushing).

A number of authors maintain, with Barnett, that the correct measure of choiceworthiness in binary decision problems is degree of ratifiability. The decision rule that instructs maximizing this quantity is, for example, an implication of such decision theories as Wedgwood 2011's *Benchmark Theory*, Podgorski 2022's *Tournament Decision*

[43] As the *Shell Game* of Skyrms 1984 and the related *Three-Option Smoking Lesion* problem presented in Egan 2007 (and credited there to Anil Gupta) illustrate, even when there exists an available ratifiable option, it sometimes seems more rational to choose an unratifiable one instead, as both EDT and CDT (in various formulations) recognize in the absence of tacked on ratifiability principles.

Theory, and Gallow 2020's *Managing the Improvement News*, which can all be viewed as attempts to generalize the graded ratifiability rule beyond binary problems.[44] The details of these generalizations are not terribly important for our purposes, though we may note that all of these theories neatly handle the cases of decision instability that originally motivated concern for CDT (e.g., they recommend randomizing in **Dicing with Death** and taking the envelope in **Frustrator**), while avoiding some of the problematic verdicts of EDT (e.g., in **Newcomb**).

If any of the above graded ratifiability theories are correct, we lose our present argument that prediction actively impairs rational deliberation. However, this is only because, like EDT, none of these theories' choice rules employ quantities that logically fix act probabilities. For example, in the binary case, the degree of ratifiability of one act A relative to another B is only a function of $U_A(A)$ and $U_A(B)$, both of which are independent of the probabilities, if such exist, of A and B. So, under graded ratifiability theories, act probabilities evade the charge of practical liability but are open to critique along more traditional lines. In particular, on these views, act probabilities fall prey to Spohn's charge of epiphenomenalism or practical impotence. So, even if I am wrong that the correct fundamental measure of choiceworthiness is given by causal expected utility and this measure is instead given by some generalization of graded ratifiability, the deliberation crowds out prediction thesis may yet remain a reasonable position.[45]

That said, I believe there are good reasons to prefer coupling CDT with **Open-mindedness** to replacing it with a graded ratifiability theory. For starters, while graded ratifiability theories yield intuitive answers regarding many puzzling decision problems, they don't seem to do so universally. In particular, I have in mind problems like **Nice Psycho-Button**.[46] Computing the degrees of ratifiability of pushing and refraining in this problem, we find that pushing is strictly less ratifiable than refraining, even though both options are ratifiable in the absolute sense and pushing seems at least intuitively like a rationally permissible (if not obligatory) course of action.[47] This puts graded ratifiability theories at odds not only with my proposal, but also with both EDT and (standard and deliberational) CDT, which, in at least some cases, recommend pushing the button in **Nice Psycho-Button**.[48] These theories are thus non-conservative in the sense of violating:

> **Conservatism:** If an act is recommended by EDT and possibly recommended by Deliberational CDT, then it is at least possibly rationally permissible.[49]

[44] Technically, as Barnett shows, Wedgewood's theory, which employs a crucial *benchmark* parameter, is only equivalent to Barnett's proposal on certain ways of setting benchmarks. However, ways of setting benchmarks other than so as to yield this agreement may involve giving up some of the purported virtues of Benchmark Theory, for example, its intuitive verdicts in cases like **Psycho-Button**.

[45] Granted, we would no longer meet Rabinowicz 2002's demand of showing that act probabilities are detrimental to rational deliberation.

[46] Bassett 2015 lodges a similar objection against Benchmark Theory, in particular.

[47] $U_{Don't}(Don't) - U_{Don't}(Push) = 0 - [(0.1)(100) + (0.9)(-250)] = 215 > 65 = [(0.9)(100) + (0.1)(-250)] - 0 = U_{Push}(Push) - U_{Push}(Don't)$.

[48] Assuming sharp act credences, CDT recommends pushing so long as one's degree of belief in pushing starts out sufficiently high.

[49] By "possibly recommended by Deliberational CDT", I mean that there exists some deliberational equi-

Given that EDT and CDT are our most popular and best developed theories of rational choice, built upon substantially different measures of choiceworthiness, **Conservatism** plausibly takes their agreement regarding the (potential) rational permissibility of an act to be a reliable indicator of the act's (potential) rational permissibility. EDT and Deliberational CDT themselves obviously satisfy this condition trivially, as do Permissive and Hierarchical CDT. But graded ratifiability theories violate **Conservatism** by uniquely singling out refraining as rational in **Nice Psycho-Button**.

A further reason to prefer the approach to rational choice sketched here to those of Barnett, Wedgwood, Podgorski, and Gallow, concerns the well-known difficulties faced by these authors when attempting to generalize the rule of maximizing degree of ratifability beyond binary decision problems to ones involving three or more options. Barnett suggests the principle that one option A ought to be preferred to another B just in case A's ratifiability relative to B is greater than B's relative to A. While this does allow us to construct a preference relation over options in more-than-binary decision problems, a preference relation so constructed is liable to include cycles, rendering the thought of maximizing according to it senseless. Recognizing this, Barnett, offers no general decision rule for many option problems, content to leave us with only constraints on preference rather than choice. That this is an unsatisfying state to leave decision theory in is attested to by Wedgwood, Podgorski, and Gallow's creative efforts to develop richer theories that offer generally applicable choice recommendations.

However, each of these valiant attempts ends up contradicting what we early on (in §2) identified as the fundamental aim of rational decision making (namely, causally promoting good outcomes) in virtue of violating **Causal Dominance**. While these theories each satisfy **Causal Dominance** in binary decision problems (they were, after all, developed in part with an eye toward satisfying this principle in **Newcomb**), this guarantee is lost in the context of decision problems involving at least three options. Clever examples due to Ahmed 2012 and Spencer and Wells 2019 suffice to bring this out in the case of each of Wedgwood, Podgorski, and Gallow's theories. Since this fact is already well known and rehearsing the proof of it would require introducing the targeted theories in more detail, I will content myself with merely registering the complaint.[50],[51]

libria relative to which the act maximizes causal expected utility, and by "possibly rationally permissible", I mean that relative to some rationally legitimate epistemic attitude the agent may adopt toward her acts, the act is rationally permissible.

[50]Wedgwood 2011 recognizes the problematic nature of violating **Causal Dominance** and tries to rescue his Benchmark Theory from this charge by supplementing it with a principle requiring that dominated options be removed prior to application of his decision rule. Briggs 2010 and Bassett 2015 have objected to the *ad hoc* nature of this move. Podgorski 2022 and Gallow 2020, meanwhile, simply concede the charge that their theories violate **Causal Dominance**.

[51]I was initially tempted to see Deliberational CDT and Graded Ratifiability's shared propensity to violate Spencer 2021's *Guaranteed Principle* as an additional reason to favor open-minded Hierarchical CDT over these alternatives. However, Sebastian Krug has convinced me to abandon this line of argument, having demonstrated in personal correspondence that Hierarchical CDT faces its own potential difficulties in this regard. Though Spencer's principle is certainly intuitive, Krug's clever examples ultimately make me doubt that the Guaranteed Principle can plausibly be seen as a fully general requirement of rationality.

7.3 General Ratifiability

There is an alternative, quite different generalization of ratifiability due to Gustafsson 2011 that goes by the name *general ratifiability*. An act A is generally ratifiable just in case there is no act B such $U_C(B) > U_C(A)$ for all available acts C. That is, an option is generally ratifiable if there is no alternative option that will foreseeably enjoy greater utility regardless of what choice is ultimately made. Causally dominated acts are never generally ratifiable, and, hence, taking only one box in **Newcomb** is neither ratifiable nor generally ratifiable. However, in other decision problems, many options that are not ratifiable are nonetheless still generally ratifiable (e.g., pushing the button in **Psycho-Button**). In finite choice problems, Gustafsson shows that generally ratifiable options, unlike ratifiable ones, are always guaranteed to exist, without invoking mixed strategies.

This enables Gustafsson to suggests that, as a first approximation, the correct theory of rational choice is given by:

> **General Ratifiability:** Rational agents deliberating over a finite action set $\mathscr{A}$ ought to choose any V-maximal member of the set of generally ratifiable acts in $\mathscr{A}$.

This theory yields all the intuitive verdicts in the decision problems so far noted and satisfies **Causal Dominance** and **Conservatism**. In fact, I think Gustafsson's theory (with the below qualifications), unlike EDT or theories of graded ratifiability, constitutes an extensionally correct account of rational choice. This is so, I suggest, because it actually agrees with Hierarchical CDT in the case of open-minded agents! An act is generally ratifiable just in case an open-minded agent would not deem it determinantly inferior to another with respect to causal efficacy. Hence, for open-minded agents, maximization of desirability amongst generally ratifiable acts is nothing other than maximization of desirability amongst $\succ$-maximal acts. Gustaffson and I have arrived at the same destination by different routes.

I am thus inclined to view the relationship of Gustafsson's proposal to my defense of **Open-mindedness** as symbiotic rather than antagonistic. The validity of **Open-mindedness** coupled with the recognition of U as the correct fundamental measure of choiceworthiness can underwrite an explanation as to why only generally ratifiable acts are potential candidates for rational selection, while the soundness of Hierarchical CDT accounts for the secondary role of desirability maximization in the statement of General Ratifiability. Gustafsson's proposal should thus not be seen as a competitor to Hierarchical CDT or as an escape route for those hoping to reject **Open-mindedness**.

Gustafsson's insightful discussion of General Ratifiability does, however, helpfully reveal some previously alluded to flaws in our initial statements of both **Open-mindedness** and Hierarchical CDT that call for revision. Gustafsson cites Arntzenius as providing the following counterexample to General Ratifiability:

> **Three Boxes:** A superintelligence offers you your pick of three boxes: A, B, and C. If the superintelligence predicted that you would opt for A, they placed two dollars in A and one dollar in B. If the superintelligence

> predicted that you would opt for B, they placed four dollars in A and three dollars in B. If the superintelligence predicted that you would opt for C, they placed one dollar in A and two dollars in B. C has been left empty regardless.

Let us once more assume that you take the predictor to be highly reliable, value money linearly, etc. Given these assumptions, both A and B qualify as generally ratifiable and hence according to General Ratifiability, considerations of desirability ought to be employed in deciding between them. But $V(B) > V(A)$, assuming a sufficiently reliable predictor, and hence General Ratifiability ends up recommending B. Similarly, if an open-minded agent faces **Three Boxes**, she will not be able to conclude that either A or B has determinately superior utility and, hence, Hierarchical CDT will recommend deciding in favor of B on grounds of auspiciousness. But this seems wrong. The only way B could possibly yield a better return than A would be if the superintelligence predicted that you would take the empty box C, which is a senseless (i.e., generally unratifiable, causally dominated) option. Given that you won't take such a senseless course of action, and should be able to recognize as much, choosing B appears inferior to choosing A.

Gustafsson suggests modifying General Ratifiability to account for cases like this by introducing a recursive hierarchy of general ratifiability notions. Say that an act A is *generally ratifiable*$_0$ just in case it is generally ratifiable and say that A is *generally ratifiable*$_n$ just in case it is generally ratifiable with respect to the subset of available options that are *generally ratifiable*$_{n-1}$. We can use this hierarchy to define the *generally ratifiable*$_*$ acts as acts that are generally ratifiable$_n$, for all natural numbers n. Gustafsson then proposes that rational choice is ultimately determined by:

> **Iterated General Ratifiability:** Rational agents deliberating over a finite action set $\mathscr{A}$ ought to choose any V-maximal member of the set of generally ratifiable$_*$ acts in $\mathscr{A}$.

This solves the **Three Boxes** problem, since, even though B is V-maximal, only A is generally ratifiable$_*$. Fortunately, we can extend the open-minded causalist story that undergirded General Ratifiability into a story that can undergird General Ratifiability$_*$ as well if we allow for suitable modifications of **Open-mindedness** and Hierarchical CDT.

Beginning deliberation as an open-minded causalist, you should be able to recognize that C is determinately causally inferior to its rivals. However, upon updating on this information, A can then be recognized as having determinately greater utility than B. This suggests that perhaps **Open-mindedness** was stated a bit too strongly: it is only at the *start* of practical deliberation that a rational agent should adopt maximally imprecise credences. However, once she can see, prior to deliberation's close, that an act is certainly unchoiceworthy relative to the fundamental standard of choiceworthiness (i.e., causal efficacy), it is admissible to update on this information and factor it into her subsequent deliberation. To make this precise, suppose that an agent with credences given by representor $\mathscr{P}$ confronts a choice set $\mathscr{A}$. Say that $a \succ_0 b$, where $a, b \in \mathscr{A}$, just in case $U_P(a) > U_P(b)$ for all $P \in \mathscr{P}$, and let $A_n \subset \mathscr{A}$ be the disjunction

of options that are *not* maximal with respect to $\succ_n$. Now define $a \succ_{n+1} b$ to hold just in case $U_P(a) > U_P(b)$ for all $P \in \mathscr{P}(\cdot | \vee_{i=0}^{n} A_i^C)$.[52] Finally, say that $a \succ_* b$ just in case $a \succ_n b$ for all natural numbers n.

These new definitions should leads us to restate Hierarchical CDT as:

> **Hierarchical CDT$_*$:** Rational agents deliberating over a finite action set $\mathscr{A}$ ought to choose any V-maximal member of the set of $\succ_*$-maximal acts in $\mathscr{A}$.

Since, given **Open-mindedness**, the $\succ_*$-maximal acts are simply those that are generally ratifiable$_*$, this statement of Hierarchical CDT$_*$ again allows us to view the theory proposed here as ultimately in harmony with Gustafsson's.

To reiterate, from the vantage point we have now reached, **Open-mindedness** is best seen as a constrain upon an agent's credences at the *start* of her deliberations. Once an agent realizes that her practical reasons (i.e., her utility judgments) tell decisively against a given course of action, A, there is nothing wrong with the agent contracting her credence in A down to zero. Indeed, this is exactly what a rational agent ought to do. In light of such a realization of determinate suboptimality, deliberation has already done its work with respect to A and is no longer in any danger of being inappropriately crowded out or impaired by prediction. But adoption of sharp act credences prior to such a conclusion remains a dangerous game that can significantly hinder the practical deliberation of otherwise rational agents, as illustrated by cases like **Dicing with Death**, **Frustrator**, etc.

8 Conclusion

It has sometimes been suggested that there is a tension between CDT and the thesis that deliberation crowds out prediction.[53] Against this trend, I have suggested that we instead see these viewpoints as friends rather than foes. Adopting a postulate like **Open-mindedness** allows us to maintain that CDT correctly characterizes the fundamental criterion of rational decision making in terms of U-maximization, while escaping the conclusion that this inevitably leads to extremely counterintuitive verdicts in contexts of decision instability. For example, having taken the deliberation crowds out prediction thesis on board, causalists are no longer committed to seeing options like flipping the coin in **Dicing with Death** as determinately inferior, in a causalist sense, to the other available options. Rather, open-minded causalists can say what seems right in such cases: that the fundamental criterion of U-maximization is unable to rule out any of the available options as irrational.

As we have seen, the possibility for considerations of causal utility to result in such evaluative indeterminacy naturally invites the question as to whether secondary considerations (e.g., desirability, graded ratifiability, etc.) might be operative in distinguishing

[52] $\mathscr{P}(\cdot | \vee_{i=0}^{n} A_i^C)$ is simply the credal representor formed by updating each member of $\mathscr{P}$ by $\vee_{i=0}^{n} A_i^C$.

[53] Levi 2000 is most explicit about this: "Neither evidential decision theorists nor causal decision theorists appreciate that deliberation crowds out prediction." (Levi 2000, p. 402)

rational choices when utility values are imprecise. Hierarchical CDT$_*$ ventures an affirmative answer to this question and suggests that it is desirability considerations that should plausibly play this supporting role. The resultant picture of rational choice is thus two-tiered: (i) first, employ causal considerations to rule out determinately causally inferior options in an iterated fashion,[54] and (ii) then amongst the remaining options choose according to what you would like to learn your choice dispositons are, i.e. maximize desirability. Following this procedure puts agents in harmony with a range of plausible rationality postulates, like **Causal Dominance** and **Conservatism**, in addition to securing intuitive verdicts in each of the decision problems we have considered, from **Newcomb** on down to **Three Boxes**.

To be sure, there remain open routes around the conclusions reached here. One could deny that rational choice is fundamentally concerned with causing good outcomes, perhaps opting to side instead with EDT in problems like **Newcomb**. Or one could insist, along with proponents of standard formulations of Deliberational CDT, that our intuitions regarding odd cases of decision instability, like **Psycho-Button** and **Dicing with Death**, are not to be trusted. Alternatively, one could try to devise new measures of choiceworthiness altogether, as the advocates of graded ratifiability have done. But I hope I have at least succeeded in making the case that each of these routes carries significant costs that can be substantially avoided by simply banning sharp act credences on the straightforward grounds that prediction impairs deliberation.

Works Cited

Ahmed, Arif. "Dicing with Death". *Analysis*, vol. 74(4), 2014, pp. 587–92.

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

———. "Push the Button". *Philosophy of Science*, vol. 79(3), 2012, pp. 586–95.

Armendt, Brad. "Causal Decision Theory and Decision Instability". *Journal of Philosophy*, vol. 116(5), 2019, pp. 263–77.

Arntzenius, Frank. "No Regrets, or: Edith Piaf Revamps Decision Theory". *Erkenntnis*, vol. 68, 2008, pp. 277–97.

Bales, Adam. "Intentions and Instability: A Defence of Causal Decision Theory". *Philosophical Studies*, vol. 177, 2020, pp. 793–804.

Barnett, David James. "Graded Ratifiability". *Journal of Philosophy*, (forthcoming).

Bassett, Robert. "A Crituque of Benchmark Theory". *Synthese*, vol. 192, 2015, pp. 241–67.

Bradley, Seamus. "Imprecise Probabilities". *The Stanford Encyclopedia of Philosophy*, 2019.

Bratman, Michael. *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

———. "Time, Rationality, and Self-Governance". *Philosophical Issues*, vol. 22, 2012, pp. 73–88.

Briggs, Rachael. "Decision-Theoretic Paradoxes as Voting Paradoxes". *Philosophical Review*, vol. 119, 2010, pp. 1–30.

[54] Or, as Gustafsson 2011 would put it, eliminate all options that are not generally ratifiable$_*$.

Egan, Andy. “Some Counterexamples to Causal Decision Theory”. *Philosophical Review*, vol. 116, 2007, pp. 93–114.

Gallow, Dmitri. “The Causal Decision Theorist’s Guide to Managing the News”. *Journal of Philosophy*, vol. 117(3), 2020, pp. 117–49.

Gibbard, Allan, and William Harper. “Counterfactuals and Two Kinds of Expected Utility”. *Foundations and Applications of Decision Theory*, ed. by C. Hooker et al., Dordrecht, 1978, pp. 125–62.

Gilboa, Itzhak. “Can Free Choice Be Known?” *The Logic of Strategy*, ed. by R. Jeffrey C. Bicchieri and B. Skyrms, Oxford University Press, 1999, pp. 163–74.

Gilboa, Itzhak, and David Schmeidler. “Maxmin expected utility with non-unique prior”. *Journal of Mathematical Economics*, vol. 18(2), 1989, pp. 141–53.

Gustafsson, Johan. “A Note in Defence of Ratificationism”. *Erkenntnis*, vol. 75(1), 2011, pp. 147–50.

Gärdenfors, Peter, and Nils-Eric Sahlin. “Unreliable probabilities, risk taking and decision making”. *Synthese*, vol. 53(3), 1982, pp. 361–86.

Hajek, Alan. “Deliberation Welcomes Prediction”. *Episteme*, vol. 13(4), 2016, pp. 507–28.

Hare, Caspar, and Brian Hedden. “Self-Reinforcing and Self-Frustrating Decisions”. *Nous*, vol. 50(3), 2016, pp. 604–28.

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

Jeffrey, Richard. *The Logic of Decision*. University of Chicago Press, 1965/1983.

Joyce, James. “Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems”. *Newcomb’s Problem*, ed. by Arif Ahmed, Cambridge University Press, 2018, pp. 138–59.

———. “Levi on Causal Decision Theory and the Possibility of Predicting One’s Own Actions”. *Philosophical Studies*, vol. 110(1), 2002, pp. 69–102.

———. “Regret and Instability in Causal Decision Theory”. *Synthese*, vol. 187, 2012, pp. 123–45.

———. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

Lauro, Greg, and Simon Huttegger. “Structural Stability in Causal Decision Theory”. *Erkenntnis*, 2020, pp. 1–19.

Levi, Isaac. “Deliberation Does Crowd Out Prediction”. *Hommage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, ed. by J. Josefsson and D. Egonssson, 2007.

———. “Rationality, Prediction, and Autonomous Choice”. *Canadian Journal of Philosophy*, vol. 23(1), 1993, pp. 339–63.

———. “Review Essay: *The Foundations of Causal Decision Theory*”. *Journal of Philosophy*, vol. 97(7), 2000, pp. 387–402.

Lewis, David. “A Subjectivist’s Guide to Objective Chance”. *Studies in Inductive Logic and Probability, Volume II*, University of California Press, 1980, pp. 263–93.

———. “Causal Decision Theory”. *Australasian Journal of Philosophy*, vol. 59, 1981, pp. 5–30.

———. “Prisoner’s Dilemma is a Newcomb Problem”. *Philosophy and Public Affairs*, vol. 8(3), 1979, pp. 235–40.

Mahtani, Anna. “Imprecise Probabilities”. *The Open Handbook of Formal Epistemology*, ed. by Richard Pettigrew and Jonathan Weisberg, 2019, pp. 107–30.
Nozick, Robert. “Newcomb’s Problem and Two Principles of Choice”. *Essays in Honor of Carl G. Hempel*, ed. by Nicholas Rescher, Dordrecht, 1969, pp. 107–33.
Podgorski, Abelard. “Tournament Decision Theory”. *Nous*, 2022.
Price, Huw. “The Effective Indexical”. *MS*, 2007.
Rabinowicz, Wlodek. “Does practical deliberation crowd out self-prediction?” *Erkenntnis*, vol. 57, 2002, pp. 91–122.
Richter, Reed. “Rationality Revisited”. *Australasian Journal of Philosophy*, vol. 62(4):392-403, 1984, pp. 392–403.
Savage, Leonard. *The Foundations of Statistics*. Dover, 1954/1972.
Skyrms, Brian. *Causal Necessity*. Yale University Press, 1980.
———. *Pragmatics and Empiricism*. Yale University Press, 1984.
———. *The Dynamics of Rational Deliberation*. Harvard University Press, 1990.
Spencer, Jack. “An Argument Against Causal Decision Theory”. *Analysis*, vol. 81 (1), 2021, pp. 52–61.
Spencer, Jack, and Ian Wells. “Why Take Both Boxes?” *Philosophy and Phenomenological Research*, vol. 99, 2019, pp. 27–48.
Spohn, Wolfgang. *Reflexive Rationality: Rethinking Decision and Game Theory*. (unpublished).
———. “Reversing 30 years of discussion: why causal decision theorists should one-box”. *Synthese*, vol. 187, 2012, pp. 95–122.
———. “Where Luce and Krantz do really generalize Savage’s decision model”. *Erkenntnis*, vol. 11, 1977, pp. 113–34.
Stern, Reuben. “Diagnosing Newcomb’s Problem with Causal Graphs”. *Newcomb’s Problem*, ed. by Arif Ahmed, Cambridge University Press, 2018, pp. 201–20.
Weatherson, Brian. “Indecisive Decision Theory”. MS.
Wedgwood, Ralph. “Gandalf’s Solution to the Newcomb Problem”. *Synthese*, vol. 14, 2011, pp. 1–33.
Williamson, Timothy Luke. “Causal Decision Theory is Safe from Psychopaths”. *Erkenntnis*, vol. 86 (3), 2021, pp. 665–85.