Normative Decision Theory

EDWARD J. R. ELLIOTT

1. Introduction

Decision theory is the interdisciplinary study of choice. Given two or more incompatible options—say, between finishing off a pile of marking or trying out a new karaoke place with colleagues—how does one choose between them? Much of the work done within decision theory concerns descriptive issues relating to human decision making; for example, what patterns exist in our decision-making behaviour, what are the psychological mechanisms behind those patterns, and how might our future choices be most accurately predicted? The majority of contemporary research on decision theory by philosophers, however, concerns normative questions: essentially, how *should* we make our choices?

This review will briefly introduce some of the major debates within normative decision theory over the past decade or so. I'll stay focused on topics that are directly concerned with how we ought to make decisions, where the 'ought' in question is subjective and pragmatic in nature. As such, I won't discuss recent applications of decision-theoretic ideas and structures to epistemic or moral issues. Despite this, the amount of new philosophical work on decision theory is vast and highly varied, so the reader should not assume that the review is complete in all respects.[1]

There are four main sections. After providing some background in §2, in §3 I will discuss the ongoing debate between causalist and evidentialist versions of expected utility theory. In §4 I'll discuss an orthogonal debate regarding attitudes towards risk. And finally, in §5 I'll discuss a number of issues concerning the need to relax the idealising assumptions standardly made in decision theory.

2. Background

Decision theory concerns how agents ought to decide when faced with some *decision problem*. We can think of a decision problem as consisting in a set of *acts*, each of which is within the agent's power to choose. These acts are mutually exclusive and jointly exhaustive: the agent must choose exactly one of them. And each act might have a range of different *outcomes*, depending on which *state of the world* is actual.

According to expected utility theory (EUT), the decision-making agent will assign to each possible outcome a subjective value (or *utility*) that reflects the strength of her preference for that outcome obtaining. Furthermore, she will typ-

[1] Particularly noteworthy omissions include decision theory and transformative experience (see especially Paul 2014), expectation gaps and the Pasadena game (see especially Hájek 2014), and a variety of developments on the axiomatic foundations of decision theory. These topics could not be given their due here without detracting substantially from the issues already discussed.

ically be uncertain as to which state of the world is actual. The agent wants to attain the subjectively best outcome in any decision problem she might happen to find herself in. However, because she's uncertain about the actual state of the world, she will likewise be uncertain about which of the acts available to her will in fact maximise her utility. The best she can do is maximise her *expected utility*. Causal decision theory (CDT) and evidential decision theory (EDT) can be understood as two different ways of precisifying this idea—they correspond to two distinct ways of defining the 'expected' in 'expected utility theory'.

Before we discuss CDT and EDT in more detail, I'll need to introduce some formalities. From here on, I'll use 'α' to refer to the decision-maker, and I'll say:

$P \succ Q$ iff α *prefers* P to Q
$P \sim Q$ iff α is *indifferent* between P and Q
$P \succcurlyeq Q$ iff $P \succ Q$ or $P \sim Q$

Preference should be understood as a kind of comparative conative propositional attitude—essentially, *wants more* or *desires more strongly*. As a rough heuristic, you might read $P \succ Q$ as saying that α would be happier to learn that P than she would be to learn that Q.

Next, the acts. It is common to describe acts as things an agent might *do* under some intentional description. However, most philosophers today follow Jeffrey (1983) in treating acts as *propositions*—usually, propositions about what the agent does. I'll use these two ways of talking about acts interchangeably. We will let $A_1, A_2, \ldots, A_n$ designate the available acts; $O_1, O_2, \ldots, O_n$ their possible outcomes; and $S_1, S_2, \ldots, S_n$ the relevant states of the world. To simplify matters, I'll assume throughout that there are only finitely many acts, states, and outcomes.

We'll also assume that outcomes are maximally specific with respect to what α values: for each outcome O and any doxastically possible P that entails O, $P \sim O$. And we'll assume that the states are what Lewis (1981) calls a *dependency hypotheses*: conjunctions of counterfactual conditionals that specify, for each act under consideration, the outcome that would result if that act were chosen. For example,

$$S_1 = (A_1 \mathbin{\square\!\!\rightarrow} O_1)\&(A_2 \mathbin{\square\!\!\rightarrow} O_2)\&\ldots\&(A_n \mathbin{\square\!\!\rightarrow} O_n)$$

Consequently, every doxastically possible act-state conjunction $A\&S$ determines a specific outcome O, where $(A\&S) \sim O$. We can represent the relationship between acts, states and outcomes using a *decision matrix*, like so:

	S_1	S_2	…	S_n
A_1	O_1	O_3	…	$O_{n'}$
A_2	O_2	O_4	…	$O_{m'}$
…	…	…	…	…
A_n	O_n	O_m	…	O_k

We'll let u be α's *utility function*; this assigns real numbers to propositions in such a way as to represent α's preferences over them on at least an interval scale. This implies, amongst other things, at least the following minimal condition:

$$P \succcurlyeq Q \text{ iff } u(P) \geq u(Q)$$

And finally, we let c be α's *credence function*, representing her degrees of belief (or credences). We assume that c satisfies the usual axioms of probability.

Given all that, EDT can be understood as saying that an act A is permissible for α just in case, of all the acts available to her, A has maximal *evidentially expected utility*, which we can define as:

$$\mathcal{V}_{\text{EDT}}(A) = \sum_S c(S_i|A)u(A\&S_i)$$

So, α ought to choose the act (or one of the acts) which maximises the utility of the outcomes she can expect, were she to conditionalise on having chosen that act. By contrast, CDT can be understood as saying that an act A is permissible for α just in case A has maximal *causally expected utility*:

$$\mathcal{V}_{\text{CDT}}(A) = \sum_S c(S_i)u(A\&S_i)$$

So, α ought to choose the act (or one of the acts) which, by her current *unconditional* credences in the various dependency hypotheses, is most likely to have the best results.[2]

3. Causal versus evidential decision theory

In most cases, there's no difference between $\mathcal{V}_{\text{EDT}}$ and $\mathcal{V}_{\text{CDT}}$, because the choice of act will be evidentially independent of which state is actual (i.e., $c(S_i|A) = c(S_i)$). But it is possible to devise situations where they come apart (or at least appear to come apart). Historically, most of the debate between EDT and CDT has centred on one type of case: *Newcomb's Problem* (Nozick 1969). Particularly telling are the 'medicalised' versions; e.g.,

> *Smoking Lesion*
> α's acts are to smoke or not smoke. She knows that smoking is highly correlated with lung cancer, but only because of a common cause—a lesion which tends to cause both smoking and cancer. Once the presence or absence of this lesion has been fixed, there's no additional correlation between smoking and cancer. She prefers smoking to not smoking, independent of whether she gets cancer, though she also strongly prefers not to have cancer.

The widespread opinion is that α should smoke, because smoking *dominates*: she's better off smoking regardless of whether she has the lesion or not, and her

[2] There are multiple ways of characterising EDT and CDT, and each should be thought of more as a cluster or related views than a single precise thing. For our purposes it's more important to note the difference between EDT and CDT than it is to note the differences between (say) alternative ways of fleshing out CDT.

smoking makes no difference to whether she has the lesion. CDT recommends smoking, whereas EDT seems to recommend against it: conditionalising on the choice to smoke makes it much more likely that α is in a state where both of her options have relatively low utility, whereas conditionalising on the choice to abstain makes it more likely that both her options will have a relatively high utility. Thus, the *Smoking Lesion* is generally considered to create troubles for EDT.

Over the past decade, however, attention has shifted somewhat from variations on *Newcomb's Problem* to another kind of case where CDT and EDT seem to give conflicting recommendations, which seem to tell against CDT in particular.[3] The following was made especially prominent by Egan (2007):

> *Psychopath Button*
> α's can choose to either press or not press the "kill all psychopaths" button. She's very confident that she's not a psychopath, and it would, she thinks, be much better to live in a world with no psychopaths. On the other hand, α is also confident that only a psychopath would press the button, and she strongly prefers living in a world *with* psychopaths to dying. Should α press the button?

Many commentators report the strong intuition that α should not press the button, and that seems to be the more common response. (It is not the universal response; see, e.g., Ahmed 2012: 387.)

According to EDT, α should not press the button. Pressing would constitute strong evidence that she is a psychopath, and hence that she's in a state where pressing will lead to the worst outcome (her own death); whereas not pressing in all states merely leads to the much less bad outcome of living in a world with psychopaths. According to CDT—at least on the way I've framed it here—α should press the button. She's confident she's not a psychopath, so by that measure pushing is most likely to lead to the best results. Given the (apparently) widespread anti-pushing intuition, the *Psychopath Button* is often touted as a counterexample to CDT.

A wide range of responses on behalf of CDT have been suggested. Cantwell (2010) argues that the informal description of the decision problem on which the anti-pushing intuition is grounded is inconsistent with the description under which CDT actually recommends pushing the button. Bales (forthcoming) argues that if we think of α's 'acts' as the intentions she might form, and combine that with a Bratmanian view of intentions, then CDT will recommend not pushing.[4] Ahmed (2012, 2014a: 61-65) argues that the basis for the anti-pushing intuition is incompatible with CDT's preferred response to *Newcomb's Problem*,

[3] *Newcomb's Problem* is still frequently discussed, of course. See, *inter alia*, (Ahmed 2014a), (Hare and Hedden 2016), (Bales 2018a), (Spencer and Wells forthcoming), (Wells forthcoming).

[4] The nature of *acts* in decision theory—e.g., whether they should be thought of as non-intentionally characterised *bodily movements*, intentionally characterised *external actions*, or internal *mental acts* like the forming of intentions or decisions—has not received much attention in recent years, though cf. (Hedden 2012).

unless we also want to accept that there are some (relatively simple and straightforward) decision problems where there are *no* rationally permissible acts. And Joyce (2012) argues that CDT, properly understood, gets the case exactly right, for reasons that I'll describe in a moment.

Several philosophers have taken the *Psychopath Button* (and similar) to show that CDT is unsatisfactory, and have offered modified—or even wholly new—theories to replace it. Arntzenius (2008) suggests we use what he calls deliberational causal decision theory (DCDT), a modified version of CDT based on a 'No Regrets' principle: agents should not be able to foresee that they will regret their decisions.[5] CDT seems to violates this principle in the *Psychopath Button*: upon choosing to push the button, and updating her beliefs on having made that choice, α should come to believe that not pushing would have been more likely to lead to better results. If she then changes her mind, and decides not to push after all, *then* she'll come to believe that pushing would have been the better option. We might imagine the committed causal decision theorist flip-flopping between pushing and not pushing until finally a choice is forced at random. So, according to Arntzenius, CDT can't help but violate the 'No Regrets' principle in a decision problem such as this.

Borrowing an account of deliberation from Skyrms (1990), Arntzenius proposes to modify CDT by invoking *mixed decisions*. A mixed decision can be modelled as a probability distribution p over the space of acts—say, $p(\text{Push}) = 0.6$, $p(\neg\text{Push}) = 0.4$—where this might at first pass be taken to represent a decision to act 'randomly', with a probability $p(A)$ of performing A (but cf. Arntzenius 2008: 292, and Ahmed 2014a: 69-70, for discussion on the intended interpretation). What's most important is that upon having 'chosen' the mixed decision associated with the distribution p, α's credence that she'll perform A should equal $p(A)$. With this as background, we are to imagine that α begins her deliberative process with credences not only about whether she is a psychopath, but also about whether she will perform this or that act. She determines the causally expected utility of her available acts on this basis, but *does not yet* make a choice. Instead, when she finds that $\mathcal{V}_{\text{CDT}}(\text{Push}) > \mathcal{V}_{\text{CDT}}(\neg\text{Push})$, she merely *raises her credence* that she'll push. This provides evidence that she's a psychopath, so she updates her credences and does the expected utility calculations anew—perhaps this time finding that $\mathcal{V}_{\text{CDT}}(\text{Push}) < \mathcal{V}_{\text{CDT}}(\neg\text{Push})$. This process is repeated, raising and lowering $c(\text{Push})$ and $c(\neg\text{Push})$, until finally an equilibrium is reached—a stable point at which further expected utility calculations no longer shift her credences towards the available acts either up or down. The equilibrium corresponds to the mixed decision that α should 'choose'.

Arntzenius' proposal has met with some resistance. Plommer (2016) argues that it requires us to calculate expected utilities in a subtly inappropriate way. Ahmed (2014b) suggests a variation on Gibbard and Harper's (1978) *Death in Damascus* case that he argues DCDT gets wrong. (See also Ahmed 2014a: 69-

[5] See (Wallace 2010) for a view that's very similar to Arntzenius', and (Gustafsson 2011) for an alternative take on the *Psychopath Button* also based upon a 'No Regrets' principle.

73.) And Joyce (2012) argues that Arntzenius' 'No Regrets' principle is acceptable only to the extent that it's already implied by CDT, and not otherwise.

With that said, Joyce's response to the *Psychopath Button* is at a glance very close to Arntzenius' own. According to Joyce, CDT properly characterised should include the constraint that α is only to make a choice on the basis of her assessment of the expected utilities *once all relevant information is in*. Since introspective evidence about one's own expected utilities counts as relevant information (and since it is presumably available to the decision-maker), Joyce argues that CDT already forces a deliberational process much like the one Arntzenius describes, and ultimately prescribes being *indifferent* between pushing and not pushing.[6]

A wholly distinct response to the *Psychopath Button* is Wedgwood's (2013) benchmark theory (BT). Like CDT, BT is designed to be sensitive especially to the counterfactual consequences of the available acts, but unlike CDT (and like EDT), it's also designed to be sensitive to the evidence that those acts provide about the causal structure of the world. The foundational idea is that the merits of an act in a state should be evaluated relative only to how well other acts do *at that state*, not how well they might have done in other states. Thus, we define for each state S_i a 'benchmark' value, b_i, and define the comparative utility of an action's outcome at a state relative to that benchmark:

$$cu(A,S_i) = u(A\&S_i) - b_i.$$

According to BT, α should choose the act with maximal expected *comparative* utility, with the expectations being determined by her credences for each state conditional on the act being chosen:

$$\mathcal{V}_{BT}(A) = \sum_S c(S_i|A)cu(A,S_i)$$

Wedgwood argues that BT gives the intuitively correct results for the *Psychopath Button* case. For critiques of the theory, see (Briggs 2010: 15-17) and (Bassett 2015).

One can see BT as a kind of hybridisation of CDT and EDT, intended to accommodate intuitions that in some hypothetical cases apparently tell in favour of CDT, while in other cases in favour of EDT. Alternative approaches to accommodating these mixed intuitions include Price's (2012) blending of CDT and EDT, at least in cases where agents have foreknowledge of events that occur by chance; MacAskill's (2016) meta-decision theory, which builds uncertainty about the correct theory of decision-making into the decision rule itself; and Bales' (2018b) decision-theoretic pluralism, according to which the concept of *permissible choice* admits of indeterminacy, with EDT and CDT corresponding to different precisifications thereof. There is also the 'no theory' theory—for instance, Briggs (2010) applies Arrow's (1950) classic impossibility result

[6] It's worth flagging that Arntzenius' and Joyce's responses to the *Psychopath Button* both presuppose that it's *possible* for α to assign credences to acts presently under her deliberation. Joyce (2002) has argued for this in earlier work, and the topic has received a small amount of attention in the recent literature: see (Hajek 2016), (Elliott 2017a), and (Liu and Price forthcoming).

for preference aggregation to argue that no single decision rule can adequately accommodate all the intuitive data.

Beyond the *Psychopath Button*, still more counterexamples to CDT have been raised in recent years. Hare and Hedden (2016: 615ff) put forward an enhanced *Newcomb*-like problem in which, they argue, CDT (and DCDT) will lead self-aware decision-makers to choose in a clearly self-destructive manner. And Ahmed (2013a, 2013b, 2014a) has suggested a variety of cases aimed at taking down CDT. To take just a single example, consider:

> *Betting on the Past*
> α places high confidence in a deterministic system of laws *L*, and must choose between two bets. The first is such that she'll win \$10 if *P*, lose \$1 otherwise. The second is such that she'll win \$1 if *P*; lose \$10 otherwise. *P* is the proposition that at some point in the past the world was thus-and-so, where *P*&*L* entails α will take the second bet.

Ahmed argues that CDT recommends taking the first bet, since it that option dominates and the choice has no way of influencing the causal structure of the situation. On the other hand, Ahmed argues, α *should*—and, given a 'soft' determinism, *can*—take the second bet, which is what EDT advises.

4. Attitudes towards risk

Orthogonal to the debates between CDT and EDT (and DCDT and BT and…) is another debate concerning the appropriate way to incorporate considerations of *risk* into our normative theories of decision making. Going at least as far back as Allais (1953), we have known that there are decision problems where ordinary agents seem to have preferences that conflict with EUT in general, regardless of whether that theory is cashed out in causalist or evidentialist terms. What's more, the preferences in question don't seem obviously *irrational*.

A simple example will suffice to make the point. Imagine that α is exactly ⅓ confident that *S*, ⅔ confident that ¬*S*, and she faces a choice between two options. On the one hand (A_1), she might take a ticket in a lottery that pays out either \$100 or \$1, depending on whether *S* or ¬*S* respectively. On the other hand (A_2), she might take \$34 unconditionally. We might represent her decision problem as follows:

	S	¬*S*
A_1	\$100	\$1
A_2	\$34	\$34

If we assume that $u(\$x) = x$, then EUT implies that α ought to be perfectly indifferent between A_1 and A_2:

$$\begin{aligned} \mathcal{V}_{\text{EUT}}(A_1) &= c(S)u(\$100) + c(\neg S)u(\$1) \\ &= \tfrac{1}{3}(100) + \tfrac{2}{3}(1) = 34 \end{aligned}$$

$$\begin{aligned} \mathcal{V}_{\text{EUT}}(A_2) &= c(S)u(\$34) + c(\neg S)u(\$34) \\ &= \tfrac{1}{3}(34) + \tfrac{2}{3}(34) = 34 \end{aligned}$$

But it is not at all hard to imagine α strictly preferring A_2. Given A_1, she'd have a reasonable shot at the higher payout, but that option has a significantly lower *minimum* payout. By contrast, there's no risk of being left with just \$1 if α takes A_2—it's a *sure thing*. And there is nothing *obviously* irrational about preferring the less risky of the two options.

Proponents of EUT have responded to this kind of example in one of two ways. On the one hand, many have argued that EUT sets the correct standards of rational decision making, so any preferences other than indifference in this kind of case must be irrational. This general thought has led many descriptive decision theorists to develop a variety of *non*-expected utility theories to accurately describe and predict the 'irrational' preferences of ordinary agents. On the other hand, some proponents of EUT have also argued that some relevant aspect of the decision problem might have been mischaracterised, and that the intuitively permissible preferences for A_2 need not be irrational after all. In particular, we might need to redescribe the outcomes to better reflect how α perceives the decision problem she faces, or at least we might need to check our assumptions about the utilities she assigns to those outcomes. Most obviously, by assuming that $u(\$x) = x$, we're in effect assuming that the utility α assigns to \$34 is situated exactly one third of the way between the utilities she assigns to \$1 and \$100. This is not required by EUT, which is consistent with A_2 being preferable whenever $u(\$34) > \frac{1}{3}(u(\$100) - u(\$1))$.

Buchak (2013, 2014) takes a different approach. Building on earlier theoretical work by Quiggin (1982) and Machina and Schmeidler (1992), her new risk-weighted expected utility theory (REU) permits *rational* sensitivity to risk in cases like these, with or without the assumption that $u(\$x) = x$. To see the difference between REU and EUT, it helps to first reformulate EUT somewhat. Assume that $u(\$x) = x$, and consider again the expected utility of A_1:

$$\mathcal{V}_{\text{EUT}}(A_1) = c(S)u(\$100) + c(\neg S)u(\$1)$$

We can read this as saying that the *value of* A_1 is the value of a ⅓ chance at \$100, plus a ⅔ chance at \$1. But we can also think of it like this: if α choses A_1, then she's guaranteed to get *at least* \$1 regardless of what happens, and in the event that S is true she'll also get \$99 *more*. Say that \$1 is the *guaranteed minimum*, and \$99 is the *conditional bonus*; the utility of the latter is equal to the utility of the better outcome minus the utility of the worse outcome. Then, EUT says that value of A_1 for α is equal to

(i) her utility for the guaranteed minimum, *plus*
(ii) her utility for the conditional bonus, *weighted by* her credence in the relevant conditions obtaining.

Hence, we can rewrite the above formula as follows:

$$\begin{aligned}\mathcal{V}_{\text{EUT}}(A_1) &= u(\$1) + c(S)(u(\$100) - u(\$1)) \\ &= 1 + \tfrac{1}{3}(100 - 1) = 34\end{aligned}$$

To put that more generally, EUT dictates that rational agents will always weight conditional bonuses by their credences towards the conditions in question.

Buchak's REU denies exactly this: decision-makers' credences *matter* according to REU, but they're not the whole story. We also need to consider *attitudes towards risk*.

Formally, we're to model α's risk-attitudes using a function *r*, which transforms α's credences before they interact with her utilities to determine the overall value of the act—by either 'inflating' or 'deflating' those probabilities in accord with whether α is risk-seeking or risk-averse. More precisely, where *r* is a (strictly increasing and continuous) function from [0,1] to [0,1], with $r(0) = 0$ and $r(1) = 1$, we define the *risk-weighted expected utility* of A_1 as:

$$\mathcal{V}_{\text{REU}}(A_1) = u(\$1) + r(c(S))(u(\$100) - u(\$1))$$

Thus, the utility of the conditional bonus ($u(\$100) - u(\$1)$) is weighted *not* by α's credence towards *S* directly, but by a function of her credences that represents her risk-attitudes. In the special case where $r(c(S)) = c(S)$, then we say that α is perfectly *risk-neutral*, so REU and EUT will amount to precisely the same thing. But if α is *risk-averse*, then $r(c(S)) < c(S)$, and she will end up assigning *less* value to A_1 than would seem permissible under EUT.

It will come as no surprise that numerous objections to REU have already been put forward in the literature.[7] Thoma and Weisberg (2017) attempt to undermine the intuitive support for REU. They argue that once all the relevant details of the agents' decision problems have been spelled out *in full*, REU fails to recapture the intuitively permissible preferences of risk-averse agents in cases like the *Allais paradox*. Thoma (2019) argues that REU and EUT amount to (at least approximately) the same thing: for an agent who sees any small-stakes decision problem she's presently faced with as just one in a long series of similar choices she'll need to face over the course of her life, REU will (given some plausible assumptions) itself recommend acting as if one is risk-neutral. And Briggs (2015) and Joyce (2017) both argue that REU permits irrational decisions in cases of sequential choice, leading agents to accept dominated strategies or leaving them susceptible to Dutch Books.

Pettigrew (2015a) demonstrates that preferences for A_2 can be explained within an EUT framework—indeed, that *any* of the preferences permitted by REU can be so explained—if we're allowed to re-describe the outcomes to make them more fine-grained. In particular, we're to suppose that one and the same coarse-grained outcome (e.g., $1, or $34) might have different utilities contingent on whether it was brought about by this or that act. Agents' attitudes towards risky options can then be encoded in their utilities towards *act-outcome pairs*, while everything else about the basic EUT decision rule is left the same. Thus, we have a range of 'risk-averse' preferences over acts, which might be rationalised by assuming *either* (i) that the decision-maker is following the REU rule with the outcomes as originally described, or (ii) the she is following the EUT rule with the more fine-grained redescription of those outcomes.

Buchak (2015) objects to Pettigrew's version of the redescription strategy by noting that the latter kind of rationalisation is less constrained than the former:

[7] For responses to some of these objections, see (Buchak 2015, 2017).

supposing that the agent's preferences satisfy the conditions of her representation theorem, there is only one set of credences and one set of utilities (defined up to an interval scale) which could rationalise those preferences consistent with the REU rule, whereas there will sometimes be distinct and incompatible sets of utilities which rationalise those preferences consistent with the EUT rule. As Buchak (2015: 846) notes, this is only a problem if we think that it ought to be possible to read agents' utilities (and credences) directly off of their preferences—an idea that has been in sharp decline in recent years (cf. Meacham and Weisberg 2011, Easwaran 2014). But there is another, more general concern with the response: Buchak is able to prove the strong uniqueness result for her REU decision rule only by invoking extremely strong idealising assumptions like those used by Savage (1954) to obtain the strong uniqueness results of his own EUT representation theorem.[8] Under more realistic assumptions, there will typically be many ways of making a set of preferences fit (or approximately fit) with either the REU or EUT decision rules.

Stefansson and Bradley (forthcoming) also adopt a version of the redescription strategy as a way to capture risk-averse preferences. Adapting Jeffrey's (1983) axiomatic framework, they enrich the underlying space of propositions to include propositions about objective chance distributions over outcomes, and represent agents' risk-attitudes *via* their utilities for these chance propositions. To motivate their way of accommodating risk-aversion in a normative framework, Stefansson and Bradley argue that (i) unlike their own account, REU is unable to accommodate the seemingly rational preferences ordinary agents tend to display in the *Ellsberg Paradox* (Ellsberg 1961), and (ii) the REU model misrepresents the psychology of risk-attitudes. Buchak (2013: 80-81) raises the former issue as a potential worry for REU as well, though it's set aside to be dealt with under future developments of the theory—specifically, those which might allow for 'imprecise' credences (see §5.2 below).

5. Deidealising decision theory

It's possible to view REU as one way of *deidealising* expected utility theory. That is—and, setting aside the redescription strategy—we *could* interpret the situation described in §4 as one in which EUT implicitly presupposes that rational agents are risk-neutral, whereas we might want our theory of good decision-making to incorporate a wider range of possible risk-attitudes.

Put in these terms, REU becomes one part in a much bigger project to broaden the scope of normative decision theory by relaxing some of EUT's many idealising assumptions. For example, on the standard way of setting things up, we typically assume that α is *aware* of all the acts available to her, that she has *precise utilities* towards each of the relevant outcomes (represented by the real-valued function *u*), and that she has *precise* and *probabilistically coherent credences* towards all the relevant states (represented by the probability function

[8] For discussion on Savage's assumptions, see (Joyce 1999: 97ff).

c). These are strong assumptions by any measure, and most theorists today think that at least some of them are too strong.

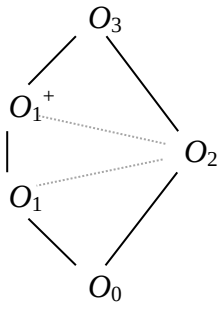
There are in fact two projects here, and philosophers have made substantial contributions to both. On the one hand, you might think that the some of the idealisations built into EUT are too demanding for agents *like us*. The rational capacities of ordinary agents like us are bounded in a variety of ways, and so we need a theory of rational decision making that we can actually live up to. This idea is the basis of the *bounded rationality* project, which I will not focus on here. (But see Weirich 2015, Elliott 2017b, and Bradley 2018, for recent work connected to this project.) On the other hand, you might think that some of the idealisations in question ask too much of even *ideally* rational beings with no special bounds on their rational capacities. In the following subsections, I will consider two different strands of this latter deidealisation project.

5.1 Incomplete preferences

Hare (2010) considers a case in which one has the option to preserve from destruction at most one of either:

> O_1: An item of significant historical value, such as the Fabergé egg
> O_2: An item of significant personal value, such as a wedding album

Even if she were ideally rational, it would not be unreasonable for α to lack any all-things-considered preference between O_1 and O_2. Moreover, α's lack of preferences might display *insensitivity to sweetening*. Let O_1^+ be just like O_1, but for the addition of a mild benefit—e.g., the Fabergé egg plus $10. If α were merely *indifferent* between O_1 and O_2, then since we can safely assume that $O_1^+ \succ O_1$, we'd also expect that $O_1^+ \succ O_2$. Yet this need not be the case: α might just as reasonably lack any all-things-considered preference between O_1^+ and O_2. Overall, then, it seems that α's preferences might permissibly take the following kind of structure (where the solid downward lines represent the asymmetric preference relations, and the dotted lines represent a symmetric *lack of preference* relation that isn't indifference):



In this case, α has *incomplete preferences* over the outcomes. It is a trivial matter to show that an incomplete preference ranking like this cannot be faithfully represented by a real-valued function *u*: since the ≥-ordering over the real num-

bers is not itself incomplete, it cannot be used to faithfully represent any incomplete $\succcurlyeq$-ordering over propositions.[9]

There has been some debate about whether the incompleteness here should be analysed as α's having *vague* preferences, or whether perhaps it highlights the need for a new kind of symmetrical preference relation that usually goes by the name 'parity' (cf. Rabinowicz 2009, Gustafsson and Espinoza 2010, and especially Chang 2014). But independent of the *source* and *nature* of the incompleteness, if incomplete preference rankings are rationally permissible then they should be incorporated somehow into our best theories of decision making.

Hare's own (weakly) preferred account of decision-making with incomplete preferences he calls *prospectism*. Let an *admissible completion* of α's preferences refer to any way of rendering her overall preference structure complete without altering any of her pre-existing preferences, and which remains consistent with the basic requirements of coherence (e.g., transitivity). For instance, in the case above there are five admissible completions corresponding to where we might place O_2 in relation to O_1 and O_1^+:

1. $O_3 \succ O_2 \succ O_1^+ \succ O_1 \succ O_0$
2. $O_3 \succ O_2 \sim O_1^+ \succ O_1 \succ O_0$
3. $O_3 \succ O_1^+ \succ O_2 \succ O_1 \succ O_0$
4. $O_3 \succ O_1^+ \succ O_2 \sim O_1 \succ O_0$
5. $O_3 \succ O_1^+ \succ O_1 \succ O_2 \succ O_0$

Each of these will correspond to different utility functions (perhaps more than one). Prospectism then says that the choice of an act is *permissible*, in general, just in case it would be permissible according to EUT under some admissible extension of α's preferences.

In the event that EUT would recommend the same act *A* under every admissible completion of α's preferences, prospectism is also obviously going to recommended *A*. Schoenfield (2014) argues that this fact generates problems for prospectism. Bales, Cohen and Handfield (2014) have also objected to Hare's proposal by giving a case where it conflicts with the following dominance-like principle that they call *competitiveness*:

> **Competitiveness:** An action *A* is permissible if, for every state, its consequences are *not worse* than the consequences of all alternative actions.

Indeed, Bales *et al* very briefly put forward a stronger version of this claim, that *A* is *obligatory* if it *and no other* actions have consequences which at every state are no worse than the consequences of any alternative actions. Their primary stated reason for finding the stronger competitiveness principle plausible is their intuition that in the following kind of case, A_1 is obligatory:

[9] Rabinowicz (2009, 2012) furthermore shows that only slightly more complicated incomplete preference structures cannot be adequately represented by *interval-valued* functions.

	S	$\neg S$
A_1	O_1^+	O_1^+
A_2	O_1	O_2

According to prospectism, both A_1 and A_2 will be permissible, but the strong competitiveness principle renders A_1 uniquely permissible. It's unclear to me, however, how widely shared Bales *et al*'s intuition about this case is. Doody (forthcoming) discusses the competitiveness principle in some depth and finds it wanting, and puts forward a weaker principle in its place—one that prospectism also violates.

Finally, Peterson (2015) has argued that if α chooses in accord with prospectism permissions, then she might be subjected to a 'weak money pump'—i.e., a sequence of acts, all of which she is *permitted* to choose, but which in combination she knows in advance are guaranteed to lose her money. In response, Kaivanto (2017) has put forward a variation on prospectism that avoids Peterson's money pumps.

5.2 Imprecise credences

A closely related strand of the deidealisation project relates to the possibility of *imprecise credences*. Imagine, for example, that before you sits an old pack of cards. You know that inside the pack some cards are missing, but you have no idea about how many nor which ones have been lost. Now compare:

> P = The global population in 2100 will be greater than 12 billion
> Q = The next card drawn from this old deck will be a heart

If you're like most people, you won't think that P is exactly as probable as Q. But is P more, or less, probable than Q—and if so, by how much exactly? You should find this hard to answer. I certainly do. Moreover, the difficulty doesn't seem to be merely that there's some fact about the strengths of my beliefs that's hard to determine, perhaps because I lack sufficient introspective evidence about my own beliefs. The problem (at least arguably) goes deeper than that: it's just not plausible that there *is* any precise value n such that P is exactly n times more (or less) probable than Q.

Joyce (2010) forcefully argues that this kind of imprecision need not be limited to non-ideal agents like us, but in fact represents the appropriate rational response to evidence which is itself often imprecise and fragmented. Along slightly different lines, Williams (2014) argues that imprecise credences are an appropriate epistemic response to indeterminate subject-matters. Precise, real-valued probability functions are not well-suited for representing credal imprecision; hence, we need to generalise our formal model of credences.

There's a wide range of models which have been proposed—see (Augustin *et al* 2014) for a recent review—but the one that philosophers typically prefer is the *credal sets* model. Rather than letting α's credences be represented by a single probability function c, on the credal sets model we use a non-empty *set* of probability functions, $\boldsymbol{C}$. Exactly how we're supposed to interpret $\boldsymbol{C}$ as a model

of α's credences is often not made fully explicit—and when it is, the intended interpretation will usually differ somewhat from person to person. Nevertheless, almost everyone agrees on at least the following two points:

1. If $c(P) = c(Q)$ for all c in $\boldsymbol{C}$, then α takes P to be as probable as Q
2. If $c(P) > c(Q)$ for all c in $\boldsymbol{C}$, then α takes P to be more probable than Q

Under certain conditions, a credal set $\boldsymbol{C}$ induces an *interval-valued* function (which we'll designate C) which summarises the range of values towards a proposition that the different functions c in $\boldsymbol{C}$ might take.[10] In practice, philosophers tend to discuss imprecise credences in terms of these intervals, though there will be cases where C leaves out some of the information contained in $\boldsymbol{C}$. (See Joyce 2010 for some examples.)

Most philosophical work relating imprecise credences to decision making has been framed in response Elga's (2010) *Two Bets* argument, aimed at showing that credences should always be precise. Imagine that α's imprecise credence for P falls within the range [0.1,0.8], and she knows she'll be offered the following two choices (in short sequence):

> *Choice 1*: Accept or Reject bet B1, <Lose \$10 if P; otherwise, win \$15>
> *Choice 2*: Accept or Reject bet B2, <Win \$15 if P; otherwise, lose \$10>

If α *accepts* both bets, she's guaranteed to make a net profit of \$5; Elga considers this sufficient reason to say that *rejecting* both bets is rationally impermissible. And yet, he argues, no *plausible* decision rule for imprecise credences gets us this result.

In saying this, Elga considers a wide range of possible decision rules—far too many to consider here. But one obvious example is worth noting: the 'permissive' rule, which is analogous to Hare's prospectism. According to this rule, an act A is permissible just in case it would be permissible according to standard EUT under *any* of the precise probability functions in $\boldsymbol{C}$.[11] If we consider each of the two choices in isolation from one another, then EUT permits rejecting B1 whenever $c(P) \geq 0.8$, and permits rejecting B2 whenever $c(P) \leq 0.4$. Consequently, Elga argues, the permissive rule permits rejecting both bets.

The most common strategy of response to Elga's argument has been to point out a number of possible decision rules for imprecise credences which, in combination with a sophisticated approach to sequential decision making, rule out rejecting both bets. Versions of this response can be found in (Sahlin and Weirich 2014), (Chandler 2014), (Bradley and Steele 2014), and (Sud 2014). A *sophisticated* decision maker knows how she is liable to choose in future decision problems, and recognises that the choices she makes now can affect which problems she's faced with in the future; thus, when deciding on B1 she will incorporate into her evaluation of the outcomes how her choice now will affect her choice regarding B2.

[10] That is, $C(P) = \{c(P) \mid c \in \boldsymbol{C}\}$; where $\boldsymbol{C}$ is convex, $C(P)$ will pick out an interval.

[11] See (Moss 2015a) for detailed discussion on this kind of view, and (Moss 2015b) for a closely related alternative.

Now consider this sophisticated approach to sequential decision-making in combination with the rule usually known as Γ-maximin, which says that an act *A* is permissible just in case its *minimal expected utility*—i.e., the lowest expected utility relative to any *c* in ***C***—is *maximal*—i.e., no less than the minimal expectations of any of *A*'s alternatives. According to Sahlin and Weirich (2014), the sophisticated follower of Γ-maximin knows, while deciding on B1, that if she takes B1 she'll also end up also taking B2, whereas if she rejects B1 then she'll also reject B2. These facts thus need to be factored into the outcomes of her present choice—in a rough sense, she's effectively choosing now whether to accept or reject both bets. Relative to any probability *c* in ***C***, the minimal expectation of accepting both is equal to the utility of winning $5, whereas the minimal expectation of rejecting both is the utility of the *status quo*; hence, α should opt to take B1 and B2.

In an erratum to his paper, Elga (2012) has agreed that sophisticated choice will help to save some decision rules for imprecise credences from his *Two Bets* argument—thought it does not save all. Mahtani (2018), however, has more recently argued that sophisticated choice will *not* help in general: the key assumption of the strategy is that α will be able to reliably predict her future choices under different suppositions about what she chooses now. But, Mahtani argues, an agent with imprecise credences will display 'unstable' betting behaviour when faced with B2, regardless of her choice regarding B1, rendering her future choice dispositions unpredictable.

Rinard (2015) takes a slightly different line in response to Elga, and in the process suggests a new supervaluationist theory of decision making for imprecise credences. According to Rinard, we should interpret *imprecise* credences as *indeterminate* credences, with the various probability functions in ***C*** seen as admissible *precisifications* of α's indeterminate credal state. Following the usual supervaluationists' line, we then say that an action is determinately permissible (or impermissible) if and only if it is permissible (impermissible) relative to every such precisification. For each bet B1 and B2, it will be indeterminate whether it's permissible to reject that bet, but rejecting *both* bets will be determinately impermissible.

In a recent paper, Bradley (forthcoming) has objected to Rinard's proposal, along with two other positive proposals put forward in Sud (2014) and Moss (2015b) both aimed at dealing with Elga's *Two Bets* argument. The essence of the objection is that these decision theories are unable to adequately explain the (apparently) rational phenomenon of *ambiguity aversion*, as exemplified in the classic *Ellsberg paradox*. Although other decision rules have been suggested which do manage to adequately deal with the *Two Bets* case and can also capture the Ellsberg preferences—such as the sophisticated Γ-maximin rule—these suffer from distinctive problems of their own. If we take the Ellsberg preferences to be normatively compelling, the upshot is that as of today, we still lack an adequate theory of decision making for imprecise credences.

References

Ahmed, A. 2012. Push the Button. *Philosophy of Science* 79: 386-395.

—— 2013a. Causal Decision Theory: A Counterexample. *Philosophical Review* 122: 289-306.

—— 2013b. Causal Decision Theory and the Fixity of the Past. *British Journal for the Philosophy of Science* 65: 665-685.

—— 2014a. *Evidence, Decision and Causality*. New York: Cambridge University Press.

—— 2014b. Dicing with Death. *Analysis* 74: 587-592.

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque. *Econometrica* 21: 503-546.

Arntzenius, F. 2008. No Regrets, or: Edith Piaf Revamps Decision Theory. *Erkenntnis* 68: 277-297.

Arrow, K.J. 1950. A Difficulty in the Concept of Social Welfare. *Journal of Political Economy* 58: 328-346.

Augustin, T., Coolen, F.P.A., de Cooman, G., and Troffaes, M.C.M. 2014. *Introduction to Imprecise Probabilities*. New York: John Wiley and Sons.

Bales, A. 2018a. Richness and rationality: causal decision theory and the WAR argument. *Synthese* 195: 259-267.

—— 2018b. Decision-Theoretic Pluralism: Causation, Evidence, and Indeterminacy. *Philosophical Quarterly* 68: 801-818.

—— Forthcoming. Intentions and Instability: A Defence of Causal Decision Theory. *Philosophical Studies*.

Bales, A., Cohen, D., and Handfield, T. 2014. Decision Theory for Agents with Incomplete Preferences. *Australasian Journal of Philosophy* 92: 453-470.

Basset, R. 2015. A critique of benchmark theory. *Synthese* 192: 241-267.

Bradley, R. 2018. *Decision Theory with a Human Face*. Cambridge: Cambridge University Press

Bradley, S. Forthcoming. A Counterexample to Three Imprecise Decision Theories. *Theoria*.

Bradley, S. and Steele, K. 2014. Should Subjective Probabilities be Sharp? *Episteme* 11 (3): 277-289

Briggs, R.A. 2010. Decision-Theoretic Paradoxes as Voting Paradoxes. *Philosophical Review* 119: 1-30.

—— 2015b. Costs of Abandoning the Sure-Thing Principle. *Canadian Journal of Philosophy* 45: 827-840.

Buchak, L. 2013. *Risk and Rationality*. Oxford: Oxford University Press.

—— 2014. Risk and Tradeoffs. *Erkenntnis* 79: 1091-1117.

—— 2015. Revisiting Risk and Rationality. *Canadian Journal of Philosophy* 45: 841-862.

—— 2017. Replies to Commentators. *Philosophical Studies* 174: 2397-2414.

Cantwell, J. 2010. On An Alleged Counter-example to Causal Decision Theory. *Synthese* 173: 127-152.

Chandler, J. 2014. Subjective Probabilities Need Not Be Sharp. *Erkenntnis* 79: 1273-1286.

Chang, R. 2014. *Making Comparisons Count*. New York: Routledge.

Doody, R. Forthcoming. Parity, Prospects, and Predominance. *Philosophical Studies*.

Easwaran, K. 2014. Decision Theory without Representation Theorems. *Philosopher's Imprint* 14: 1-30.

Egan, A. 2007. Some Counterexamples to Causal Decision Theory. *Philosophical Review* 116: 93-114.

Elga, A. 2010. Subjective Probabilities should be Sharp. *Philosopher's Imprint* 10: 1-45.

—— 2012. Errata for Subjective Probabilites should be Sharp. Published online at: http://www.princeton.edu/_adame/papers/sharp/sharp-errata.pdf.

Elliott 2017a. Probabilism, Representation Theorems, and Whether Deliberation Welcomes Prediction. *Erkenntnis* 82: 379-399.

—— 2017b. A Representation Theorem for Frequently Irrational Agents. *Journal of Philosophical Logic* 46: 467-506.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics* 75: 643-669.

Gibbard, A. and Harper, W. 1978. Counterfactuals and two Kinds of Expected Utility. In *Foundations and Applications of Decision Theory*, eds. C.A. Hooker, J.L. Leach, and E.F. McClennan. Dordrecht: Reidel.

Gustafsson, J.E. 2011. A Note in Defence of Ratificationism. *Erkenntnis* 75: 147-150.

Gustafsson, J.E. and Espinoza, N. 2010. Conflicting reasons in the small-improvement argument. *Philosophical Quarterly* 60: 754-763.

Hájek, A. Unexpected Expecations. *Mind* 123: 533-567.

—— 2016. Deliberation Welcomes Prediction. *Episteme* 13: 507-528.

Hare, C. 2010. Take the Sugar. *Analysis* 70: 237-47.

Hare, C. and Hedden, B. 2016. Self-Reinforcing and Self-Frustrating Decisions. *Nous* 50: 604-628.

Hedden, B. 2012. Options and the Subjective Ought. *Philosophical Studies* 158: 343-360.

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

Joyce, J. 1999. *The Foundations of Causal Decision Theory*. New York: Cambridge University Press

—— 2002. Levi on the causal decision theory and the possibility of predicting one's own actions. *Philosophical Studies* 110: 69-102

—— 2010. A Defence of Imprecise Credences in Inference and Decision Making. *Philosophical Perspectives* 24: 281-323.

—— 2012. Regret and Instability in Causal Decision Theory. *Synthese* 187: 123-145.

—— 2017. Commentary on Lara Buchak's Risk and Rationality. *Philosophical Studies* 174: 2385-2396.

Kaivanto, K. 2017. Ensemble Prospectism. *Theory and Decision* 83: 535-546.

Lewis, D. 1981. Causal decision theory. *Australasian Journal of Philosophy* 59: 5-30.

Liu, Y. and Price, H. Forthcoming. Heart of DARCness. *Australasian Journal of Philosophy*.

MacAskill, W. 2016. Smokers, Psychos, and Decision-Theoretic Uncertainty. *Journal of Philosophy* 113: 425-445

Machina, M.J. and Schmeidler, D. 1992. A more robust definition of subjective probability. *Econometrica* 60: 747-780.

Mahtani, A. 2018. Imprecise Probabilities and Unstable Betting Behaviour. *Nous* 52: 69-87.

Meacham, C.J.G. and Weisberg, J. 2011. Representation Theorems and the Foundations of Decision Theory. *Australasian Journal of Philosophy* 89: 641-663.

Moss, S. 2015a. Time-Slide Epistemology and Action Under Indeterminacy. In *Oxford Studies in Epistemology*, eds. T.S. Gendler and J. Hawthorne. Oxford: Oxford University Press.

—— 2015b. Credal Dilemmas. *Nous* 49: 665-683.

Nozick, R. 1969. Newcomb's Problem and Two Principles of Choice. In *Essays in Honor of Carl G. Hempel*, ed. N. Rescher. Dordrecht: Reidel.

Quiggin, J. 1982. A Theory of Anticipated Utility. *Journal of Economics Behavior & Organization* 3: 323-343.

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

Peterson, M. 2015. Prospectism and the weak money pump argument. *Theory and Decision* 78: 451-456.

Pettigrew, R. 2015a. Risk, Rationality, and Expected Utility Theory. *Canadian Journal of Philosophy* 45: 798-826.

Plommer, B. 2016. A New Problem with Mixed Decisions, Or: You'll Regret Reading This Article, But You Still Should. *Erkenntnis* 81: 349-373.

Price, H. 2012. Causation, Chance, and the Rational Significance of Supernatural Evidence. *Philosophical Review* 121: 483-538.

Rabinowicz, W. 2009. Incommensurability and Vagueness. *Aristotelian Society Supplementary Volume* 83: 71-91.

—— 2012. Value Relations Revisited. *Economics and Philosophy* 28: 133-164.

Rinard, S. 2015. A Decision Theory for Imprecise Probabilities. *Philosopher's Imprint* 15: 1-16.

Sahlin, N. and Weirich, P. 2014. Unsharp Sharpness. *Theoria* 80: 100-103.

Savage, L.J. 1954. *The Foundations of Statistics*. New York: Dover.

Schoenfield, M. Decision making in the face of parity. *Philosophical Perspectives* 28: 263-277.

Skyrms, B. 1990. *The Dynamics of Rational Deliberation*. Cambridge: Harvard University Press.

Stefansson, H.O. and Bradley, R. Forthcoming. What Is Risk Aversion? *British Journal for the Philosophy of Science*.

Sud, R. 2014. A Forward Looking Decision Rule for Imprecise Credences. *Philosophical Studies* 167: 119-139.

Thoma, J. 2019. Risk Aversion and the Long Run. *Ethics* 129: 230-253.

Thoma, J. and Weisberg, J. 2017. Risk Writ Large. *Philosophical Studies* 174: 2369-2384.

Wallace, D. 2010. Diachronic Rationality and Prediction-Based Games. *Proceedings of the Aristotelian Society* 110: 243-266.

Wedgewood, R. 2013. Gandalf's solution to the Newcomb problem. *Synthese* 190: 2643-2675

Weirich, P. 2015. *Models of Decision-Making*. Cambridge: Cambridge University Press.

Wells, I. Forthcoming. Equal Opportunity and Newcomb's Problem. *Mind*.

Williams, J.R.G. 2014. Decision-Making Under Indeterminacy. *Philosopher's Imprint* 14: 1-34.