Objects of Choice

Wolfgang Schwarz*
Draft, 3 November 2018

Abstract. Decision theory says that rational agents choose acts that maximize expected utility. But the expected utility of an act depends on how it is represented. An adequate representation of the acts available in a given choice situation should, on the one hand, contain no information about the world of which the agent is unaware. On the other hand, it should be sufficiently rich to distinguish all available acts from one another. These demands often come into conflict, so that there seems to be no adequate representation of an agent's options at all. After reviewing existing proposals for how to construe decision-theoretic options and finding them all wanting, I suggest that decision-theoretic models should include a special domain of "virtual" option propositions to serve as objects of practical deliberation.

1 Introduction

According to decision theory, rational agents choose options that maximize expected utility, relative to the agents' credences and utilities. But even if we know the credence and utility function of a rational agent in a given situation, we cannot predict what the agent will do, unless we also know what options are available: what are the alternatives whose expected utility she is meant to compare?

The question can perhaps be ignored if we take decision theory to be a formal study of abstract decision problems. But it has to be addressed if we want to treat decision theory as a psychological model – an idealized, high-level model of the connection between (graded) belief, desire, and action. Such a model may, in turn, be understood in different ways. It could be a descriptive model, purporting to explain choices made by actual humans; it could be a normative model, prescribing the choices people ought to make; or it could be a constitutive model, implicitly defining what it is to be an agent with such-and-such beliefs and desires. On each interpretation, we can't assume that the available options are simply given as part of a well-defined decision problem.[1]

So what are the options available to an agent in a given decision situation? The naive answer is that the options are simply the acts the agent can perform. I think this

* Thanks to Romy Jaster and Theodore Korzukhin for helpful discussion.

1 The problem of defining the available options also arises for many rivals to Bayesian decision theory; I will stick with Bayesian decision theory for the sake of concreteness and familiarity.

essentially right. But the acts must be represented in a suitable way. If you don't know that the lottery ticket you are buying is a losing ticket, for example, it would be wrong to represent the chosen act as an act of *buying a losing ticket*, so that its expected monetary payoff is zero.

Intuitively, an adequate representation of an available act should not reveal anything about the world of which the agent is unsure (such as whether the ticket is a loosing ticket). On the other hand, the representation should be strong enough to distinguish the act from any other act the agent could choose. As we will see, these demands can come into conflict.

Another challenge arises from the supposed "subjectivity" of decision theory. Assume an agent believes her options are A and B, while in fact they are A and C; each of B and C has greater expected utility than A. Decision theory, as a theory of subjective rationality, arguably should not predict (or prescribe) that the agent does C, given that she is not aware that C is an option. But we also don't want to predict (or prescribe) that the agent does B, which she can't do.

I will lay out these problems more carefully in sections 2 and 3, after some more stage-setting, and in the course of discussing some general constraints on an agent's options. In sections 4 and 5 I will then consider two proposals for how to construe decision-theoretic options. The first identifies options with inner acts of trying, intending, or deciding. The second construes options not as single acts, but as probability measures over a range of acts. Neither proposal, I will argue, fully gets around the above problems.

With all hopes for an easy solution dashed, I will turn to my own proposal in sections 6 and 7. In short, I will suggest that decision-theoretic models should postulate a special domain of option propositions to serve as objects of practical deliberation. Each act an agent can perform corresponds to an option proposition insofar as choosing the proposition would make the agent perform the corresponding act, but the propositions do not transparently describe or represent the acts. They do not transparently describe or represent anything.[2]

2 What we can do

Let's start with the idea that an agent's options are whatever acts she can perform. For that to be plausible, we need to clarify how 'act' and 'can' should be understood.

First, 'can'. Ability modals are notoriously polysemous. There's a sense in which I *can* play the piano even when there's no piano around. There's a sense in which I *can't* accept an invitation to dinner if I have other plans. There's a sense in which determinism

2 In some respects, the present paper is a dual to [Schwarz 2018], which tackles an analogous puzzle about perception and defends an analogous solution.

implies that we can't do anything except what we actually do. None of these senses are relevant to our present topic.

To understand the relevant sense of 'can', compare falling out of a helicopter with reaching a junction on a hike. When you fall out of a helicopter, you will fall to the ground, and there is little you can do about that. Not so when a hike leads you to a junction. Here, whether you end up turning right or left is sensitive to your psychological state: to your beliefs, desires, intentions, fears, hopes, and whims. In that sense, what you do is under your psychological control. You face a choice.

These are the kinds of choices studied in decision theory. They do not require a strong, libertarian kind of freedom. Decision theory is widely used in artificial intelligence, where it is taken for granted that the (artificial) agent's actions can be predicted from its internal architecture and the inputs it receives. If you build a robot, you don't need to specify that it should drop towards the Earth when it falls out of a helicopter; it will do that no matter what internal states and decision rules you build in. When the robot reaches a junction, by contrast, the outcome is controlled by its internal state: vary the state, and the robot will choose different paths. That is all the freedom we need.

The classical conditional analysis seems to gets things roughly right. On the conditional analysis, an agent *can* perform an act just in case she *would* perform it if she *tried* or *intended* to do so. For reasons that will become clear, I prefer a slightly weaker analysis, on which an agent can perform an act just in case there is some possible variation of the agent's motivational state that would make her perform the act.[3] I will say a little more on what should count as a 'possible variation' of a 'motivational state' in section 7; for now, we will have to make do with an informal, intuitive understanding.

Next, 'act'. Suppose one Sunday afternoon you come across a shrub of blueberries and decide to try one. So you *eat a berry from the bush*. You thereby also *eat a blueberry from the bush*, and you *eat a berry on a Sunday afternoon*. Action theorists debate whether these are three different acts or three descriptions of a single act. Without purporting to settle any dispute about the ordinary conception of acts, I will adopt the second usage: you perform a single act – an act of eating that particular blueberry on that particular day under those particular circumstances.

On this usage, we obviously can't assume that rational agents have perfect knowledge of what acts they are performing. You may not know that it is Sunday, or that the berry you're trying is a blueberry. Suppose you give some credence to the possibility that the berry is a poisonous tutsan berry. When you consider eating the berry, you should then take this possibility into account. That is, to assess the expected utility of your choice, we must somehow bracket the fact that the act you consider is an act of eating a blueberry.

How can we do this? One approach is due to Leonard Savage [1954]. Savage represents

3 The counterfactual ('... would ...'), like all counterfactuals in this paper, must be understood in a "non-backtracking" sense (compare [Lewis 1979b]).

acts by abstract functions from "states of nature" to "outcomes". Your act of eating the berry, for example, would be represented by a function that maps states in which the berry is a blueberry to the pleasant outcome of eating a blueberry, and states in which the berry is a tutsan berry to the less pleasant outcome of eating a tutsan berry. Unfortunately, Savage does not explain how the acts an agent can perform should be represented as functions from states of nature to outcomes, and it is hard to fill the gap.

An alternative approach has been suggested by Richard Jeffrey [1965, 1992].[4] Jeffrey draws no sharp line between acts on the one hand and states of nature on the other: "the human agent is taken to be part of nature and his acts are thus ingredients in states of nature" [Jeffrey 1992: 226]. Formally, states, acts, and outcomes are all represented as propositions. Your act of eating the berry, for example, could be represented as the proposition *that you eat a berry from the shrub*, which leaves open whether the berry is a blueberry or a tutsan berry.[5]

Jeffrey's framework has some advantages over Savage's, which is why it will be assumed in what follows. For example, if acts, states, and outcomes are all propositions, we can take into account logical and probabilistic connections between these elements, which is put to use not only in Jeffrey's own formulation of decision theory but also in many of its rivals, including [Gibbard and Harper 1978], [Lewis 1981], [Sobel 1986], [Skyrms 1984], and [Joyce 1999]. Without appeal to such connections, it is hard to give a satisfactory account of Newcomb Problems (as argued in [Joyce 1999: 117ff.]), or to model the way rational belief changes through deliberation (see [Skyrms 1990], [Joyce 2012]).

So we must distinguish between the informationally rich *acts* an agent can perform and the *options* whose expected utility they are supposed to maximize. Options, we'll assume, are propositions. How are the options related to the acts?

A few conditions seem obvious. First, we don't want decision theory to predict that an agent chooses an option if she is incapable of making the relevant option proposition true. In other words, each option should be a (typically incomplete) description of an act the agent can actually perform. Let's call this the *Ability condition* on an agent's options.

> **Ability.** A proposition A is an option (for an agent in a given choice situation) only if the agent can make A true, insofar as there is some possible variation of the agent's motivational state that would render A true.

4 Jeffrey actually has two accounts of options. Here I present the first; I'll discuss the second in section 5.

5 By a 'proposition', I mean the kind of thing that is in the domain of an agent's credence function. To avoid stylistic awkwardness, I will often use infinitive verb phrases ('eating a blueberry') instead of that-clauses ('that you eat a blueberry') to express candidate option propositions, but nothing important will hang on that – though see [Perry 1979] and [Lewis 1979a] for arguments that verb phrases are actually more perspicuous representations of the relevant propositions.

We should also impose a converse condition, that different acts an agent can perform should be represented as different options. If an agent is at a junction where she could turn left or right depending on her motivational state, it won't do to say that her only option is the proposition that she turns left, or the tautologous proposition that is true no matter what. This converse condition might be expressed as follows.

> **Cover.** For any proposition B the agent can make true (in the above sense) there is some option A such that if the agent were to choose A then B would be the case.

Another condition is needed to rule out a certain kind of redundancy among an agent's options. If you have the option of raising your left hand and another option of raising your right hand, then we may not want to count *raising a hand* as an additional, third option. This could be achieved by the following condition.[6]

> **Maximality.** A is an option only if there is no other option that entails A.

Together, Ability, Cover, and Maximality paint the following picture. Take an agent in a concrete decision situation. For any possible variation V of the agent's motivational state, there is a conjunction $@V = \bigwedge\{P : V \mathrel{\Box\!\!\rightarrow} P\}$ of everything that would be the case if V were actual – the *comprehensive outcome* of V, as we might call it. $@V$ may not be an individual possible world, but it will be highly specific.[7] So the comprehensive outcomes of all possible variations of an agent's motivational state form a set of disjoint, tiny regions in logical space. The agent's options are a "widening" of this set, expanding each region while keeping them disjoint.[8]

The remaining task is to say how the regions should be expanded – intuitively, how the acts an agent can perform should be described by her options.

Some descriptions are too narrow, including too much information about a given act. In the berry scenario, the proposition *that you eat a blueberry*, for example, contains too

6 The Maximality condition is often assumed in discussions of objective moral permission and obligation, see e.g. [Brown 2018]. Not all arguments for and against adopting the condition in that context carry over to the present context; see fn. 9 below.

7 One reason why $@V$ may not be an individual possible world is that agents don't have perfect, microscopic control over their actions. For example, if in fact you don't toss a coin, and V is a variation of your motivational state that would cause you to toss a coin, then $@V$ may neither settle that the coin lands tails nor that it lands heads (compare [Vessel 2003]). Similarly, one might argue that in a deterministic world, it is neither the case that the distant past would be different, nor that the laws would be different if an agent had made a different choice.

8 Lewis [1981: 308] seems to have a similar picture in mind. He suggests that an agent's options are given by the partition of logically strongest propositions which the agent "can make true at will". Lewis does not explain the intended reading of this crucial phrase, nor why we should expect the relevant propositions to form a partition.

much information. If this isn't obvious, let's look at the computation of expected utilities. In Jeffrey's "evidential" decision theory, the expected utility of an option O is defined as the credence-weighted average of the utility of any possible outcome R conditional on O:

$$EU(O) = \sum_R U(R)Cr(R/O).$$

Let R_T be the undesirable outcome of having consumed a tutsan berry. If B is the proposition *that you eat a blueberry*, then $Cr(R_T/B)$ is plausibly close to zero; so the badness of possible poisoning does not figure in $EU(B)$. If we treat B as one of your options, it might have maximal expected utility even if you give significant credence to the possibility that the berry is a tutsan berry. That is clearly wrong.

The same problem arises in "causal" decision theory, where expected utility is defined in terms of the credence given to the outcomes on the *subjunctive* supposition that the relevant option *would be* chosen (see [Joyce 1999]; the backslash indicates subjunctive supposition):

$$EU(O) = \sum_R U(R)Cr(R\backslash O).$$

Again, $Cr(R_T\backslash B)$ is plausibly around zero, so $EU(B)$ disregards the possibility of poisoning.

So the proposition *that you eat a blueberry* is too specific, providing too much information about the relevant act. Other propositions are too unspecific, providing too little information. For a simple (albeit artificial) example, suppose you are rewarded for unintentionally closing your eyes but punished for intentionally closing them. Having no means of causing yourself to unintentionally close your eyes, you rationally decide not to close your eyes. Here the most likely circumstances under which you would close your eyes might well be circumstances under which you do so unintentionally. As a consequence, *closing your eyes* has comparatively high expected utility (in both evidential and causal decision theory, as $Cr(\mathit{Reward} / \mathit{Close})$ and $Cr(\mathit{Reward} \backslash \mathit{Close})$ are both high). So if we count *closing your eyes* as an option, we would wrongly conclude that it would have been rational to choose it.[9]

We've seen that some act descriptions that are too specific, and others that are too unspecific. But what's the general rule? What information, in general, should be included in an option proposition?

9 The present example also motivates the Maximality condition: if we treat both *closing* and *intentionally closing* as options, the former might come out as maximising expected utility. The literature on Maximality mostly focuses on cases in which a logically weaker option is an (exclusive) disjunction of more specific options. This makes the issue rather subtle, as one might conjecture that a disjunctive option has maximal expected utility only if all its disjuncts do, in which case including the disjunctive option would do no harm. However, [Sobel 1983] shows that the conjecture is true only in evidential decision theory.

3 Further conditions

Here is an attractive thought that might point towards an answer. Decision theory is commonly understood to be a “subjectivist” or “internalist” theory, specifying what an agent ought to do *from her own perspective*, *in light of her beliefs and desires*. So the recommendations of decision theory should not depend on external matters unless the agent is aware of these matters.

This suggests that we we should strengthen the Ability condition. If A is an option, not only must the agent have the ability to make A true, she must also be certain that she has that ability:[10]

> **Modal Certainty.** A proposition is an option for an agent only if the agent is certain that she can make it true.

Modal Certainty would explain why *eating a blueberry* is too specific, as long as you give some credence to the hypothesis that you’re looking at a tutsan shrub: you then can’t be certain that you can eat a blueberry.

In other cases, however, Modal Certainty appears too weak to rule out unwanted option propositions. Imagine you have a choice between two berries. You know that one is a tutsan berry, the other a blueberry, but you don’t know which is which. *Eating the blueberry* then shouldn’t count as an option, even though you may well be certain that you can make it true – at least on the reading of ‘can’ I suggested in the previous section: you may be certain that there is some variation of your motivational state that would cause you to eat the blueberry.[11]

A more urgent problem with Modal Certainty is that it appears to be too strong. Whether or not an agent can make true a given option propositions is presumably a contingent matter: not all possible agents in all possible decision situations have the exact same options. But why should a rational decision-maker be absolutely certain of such contingent matters?

The problem is exacerbated by the Cover condition, which requires that each act an agent can perform is represented by a distinct option proposition. An example from [Jeffrey 1968: 37] brings out the point. Jeffrey considers an experienced marksman aiming at a distant target. The marksman’s posture, the way he holds his rifle, the

10 In print, Modal Certainty has been most prominently defended by J.H. Sobel; see [Sobel 1980: 178–80], [Sobel 1983: 199f.], [Sobel 1986: 155f.], [Sobel 1988: 8]; also [Hedden 2012: 348].

11 The problem could be avoided by stipulating that ‘can’ in Modal Certainty is to be understood, say, in accordance with the conditional analysis. But that would lead to other problems. For example, arguably an agent could have the option of *intending to travel to Paris* even if they don’t have the capacity to intend to intend to travel to Paris; Modal Certainty would implausiblly demand that the agent must nevertheless be certain that if they intended to intend to travel to Paris, they would succeed.

direction in which he is pointing, are all carefully attuned to his information about the wind conditions, the quality of the rifle, and so on. If the wind were to pick up, the marksman would adjust his posture. So the marksman has fine-grained control over how he holds his rifle. He is, in effect, choosing between thousands of torque configurations in his joints, although that's obviously not how he conceptualizes his options. To model the marksman's decision problem, we need thousands of option propositions corresponding to different postures and different ways of aiming at the target. By Modal Certainty, there must be thousands of relevant, distinct propositions of which the marksman is certain that he can make them true. What could these propositions be? And why should we assume that the marksman is absolute certainty that he can make them true?

Let's call this the *cover problem*. Another problem with Modal Certainty is the *problem of internalism*. The problem arises because Modal Certainty is not actually a strengthening of Ability: an agent can be certain that she can make something true even if in fact she can't. Suppose an agent can't fly but is certain that she can. We don't want decision theory to predict or recommend that the agent will fly. So we still need the Ability condition to rule out the flying option. But then we haven't really secured internalism: an agent's options still seem to depend on external facts of which she may be oblivious.[12]

On reflection, it is not clear how internalism was meant to support Modal Certainty in the first place. Informally, internalism demands that the available options should be "accessible" to the agent; decision theory should, in principle, be implementable by a local computational process whose output isn't directly sensitive to facts about the agent's environment, unmediated by perception or belief. But this doesn't mean that the agent must be certain about her options. Compare beliefs and desires. These surely count as "accessible" in the relevant sense. But as [Skyrms 1980] convincingly argues, a rational decision-maker need not be certain of her own beliefs and desires. Why would she have to be certain about her options?

Admittedly, some authors seem to assume that decision theory requires consciously computing the expected utilities of all one's options. This might require some kind of knowledge of the available options, and of the relevant credences and utilities. But the proposed interpretation of decision theory makes the theory completely implausible as a descriptive, normative, or constitutive model. Real people don't consciously compute expected utilities whenever they face a choice, and there is no reason to think that they

12 In response, Sobel ([1980], [1986]) stipulates that a decision-maker is never mistaken when she is certain about what she can do. That might be plausible if there are few things of which the decision-maker is certain. But as we've just seen, Modal Certainty (in the presence of Cover) requires that there must be lots of non-trivial things of which we are certain that we can make them true. Why couldn't we be mistaken about these things? Brian Hedden [2012: 352f.] offers a possible explanation; see footnote 15 below.

should, or that ideally rational people would.[13]

So I'm skeptical about Modal Certainty. A more promising alternative can be motivated by the fact that rational deliberation provides "knowledge without observation", in Elisabeth Anscombe's [1957: sec.8] memorable words: if you know that you can raise your arm, and you decide to raise it, then you know that your arm will go up; no further observations are needed. In the framework of Bayesian decision theory, the effect of deliberation on rational credence has been most thoroughly studied by Brian Skyrms (esp. [Skyrms 1990]). His models imply that deciding in favour of an option goes along with becoming certain of that option.

We can now turn this around (as [Jeffrey 1968] notes): if you couldn't become certain of a proposition merely through rational deliberation, then the proposition isn't one of your options. In the berry scenario, for example, if you're not sure whether the berry is a blueberry, you couldn't rationally become certain that you'll be eating a blueberry merely through deliberation. So we have another explanation of why *eating a blueberry* is too strong:

> **Deliberational Certainty.** A proposition A is an option for an agent only if the agent could become rationally certain of A merely through making a decision.

Does this avoid the problems for Modal Certainty? Unfortunately, no. For one thing, what guarantees that there are suitable, non-trivial propositions of which rational agents can become absolutely certain through deliberation? What are the thousands of propositions of which the marksman could rationally become certain? The cover problem remains.

The problem of internalism also remains. That's because Deliberational Certainty, like Modal Certainty, arguably can't replace the Ability condition. Suppose deliberation could make you rationally certain that you will, say, raise an arm. Why should this entail that you are in fact able to raise the arm?

We could now go through other conditions in place of Modal Certainty and Deliberational Certainty. But arguably the problems we have encountered will arise no matter how the missing condition is spelled out.

Let me recap how we got here. We have assumed that each act an agent can perform (on a particular reading of 'act' and 'can') is represented by an option proposition. To avoid incorrect estimations of expected utility, these propositions must not reveal external facts about the world of which the agent is unsure. So the propositions may have to be rather weak. On the other hand, each option proposition must distinguish the relevant act from all other acts the agent can perform. We need thousands of disjoint options

13 Here I agree with [Pettit 1991: 167–169], [Jackson 1991: 468–471], [Maher 1993: 5–8], [Joyce 1999: 80], among many others.

to model the marksman's decision problem. So the option propositions must be rather strong. These forces point in opposite direction, creating the cover problem.

The problem of internalism arises because whether or not an agent can make true a non-trivial proposition always seems to depend on external matters – for example, on the presence or absence of a Frankfurt-style agent poised to intervene as soon as she detects an inclination to make true the proposition (compare [Frankfurt 1969]). Given that something is an option only if the agent can actually make it true, it seems to follow that an agent's options always depend on external matters – clashing with the supposed internalism of decision theory.

4 Decisions

One might hope to escape the problems we have run into by construing an agent's options not as (propositions describing) overt acts, but as (propositions describing) internal events of *trying*, *intending*, or *deciding*. Proposals along these lines have been made in [Sobel 1971], [Sobel 1983], [Weirich 1983], [Joyce 1999: ch.2], and [Hedden 2012] (see also [Pollock 2002] for critical discussion).

It is easy to see why this move can seem appealing. Right now, for example, I'm pretty sure that I could, if I wanted, raise my arms. But I'm not entirely sure. There's a small chance that the muscles in my arms have just stopped working. Indeed, I'm not sure that I even have arms: a small part of my credence goes to the hypothesis that I'm an armless brain in a vat. As a consequence, *raising my arms* should not count as one of my options. For suppose I give low utility to the various scenarios in which I can't raise my arms – to scenarios in which my arm just got paralysed or in which I'm a brain in a vat. If *raising my arms* were an option, it would then have greater expected utility than not raising my arms simply because the proposition that I raise my arms rules out the undesirable scenarios of being paralysed or envatted. But surely my desire not to be paralysed or envatted gives me no reason to raise my arms. I'm not irrational if I don't raise my arms, given the stated beliefs and desires.

So raising my arms is not one of my options. Intuitively, I can *try* or *decide* or *intend* to raise my arms, but whether I succeed depends on whether I'm a brain in a vat and whether my muscles still work – things I can't simply make true or false by an act of the will.

There are other reasons to allow for intentions or decisions as options. We frequently make decisions not only about what we are going to do right now, but also about what we are going to do later. We make decisions about how to spend the afternoon, what to cook for dinner, or where to go on holiday. The immediate outcome of these decisions is not an overt act, but an intention. Indeed, sometimes the main reason for deciding to perform an act lies not in features or consequences of the act, but in features or consequences of the decision. If you are anxious about a certain choice, it may be reasonable to make a decision, just to stop worrying and calm your mind. In Gregory Kavka's [1983] toxin puzzle, the reason for deciding to drink the toxin is that the resulting intention will be rewarded. (Similarly in the more realistic scenarios on which the puzzle is modelled.)

How might construing options as decisions or intentions help with the problems of internalism and cover? Well, consider the marksman from the previous section. We have assumed that the marksman can choose among thousands of options. So he can make thousands of distinct decisions. Construing options as decisions immediately guarantees that we have a distinct option for each decision an agent could make. That deals with the cover problem. To secure internalism, we might hope that even though overt acts can always be thwarted by external forces, rational agents at least have full control over what they decide or intend.

But let's have a closer look. To begin, decisions, like overt acts, can be described in many ways, and the mode of presentation matters to their expected utility. (In the berry situation, your decision to try a berry is a *decision that causes you to eat a blueberry*, but that's not an adequate mode of presentation.) So again we have to decide how the available decisions should be represented or described by the agent's option propositions.

In ordinary language, decisions and intentions are commonly represented as propositional attitudes: an agent S decides or intends to ϕ, for some propositional content ϕ. There is some debate over whether intentions involve a *sui generis* attitude or whether they can be analysed in terms of credences and utilities (see e.g. [Bratman 1987]); either way, the cover problem returns.

For consider again our marksman. The marksman presumably intends to hit the target. Does he decide to hit it? Maybe. Or maybe he decides to shoot at it. But these propositions are far too unspecific. They don't single out the marksman's choice among the thousands of nearby alternatives. As I said earlier, there is a reason why the marksman holds his rifle the way he does, at a specific height, pointing in a specific direction. The reason has something to do with his desire to hit the target and his awareness of the wind etc. So what is the option he is choosing, construed as a propositional attitude? What is its propositional content? Does he decide *to adjust his joint torques in manner* $\vec{V}$, or *to hold his rifle at height* h *and orientation* α, β, γ?[14]

In some sense, I guess, he does. But in that sense, the marksman has probably no idea of what option he is choosing. Suppose we offered the marksman a prize if he decides to adjust his joint torques in the precise manner $\vec{V}$. Decision theory says that if A is one of your options, and you are certain that choosing A leads to a desirable outcome (and there are no better options, etc.), then you should choose A. So if we take the marksman's options to include a decision to adjust his torques in manner $\vec{V}$, we'd have to say that he

14 Orientations in 3D space are commonly specified by three angles, known as "Euler angles".

would be irrational if he didn't win the prize. Clearly that's wrong. Rationality does not require knowing the precise torque configurations or angles determined by one's choices.

So even if we construe options as decisions or intentions, we're stuck with the cover problem.

We're also stuck with the problem of internalism. Whether an agent can make a given decision still depends on all sorts of cognitively external matters – say, on whether she has enough energy or willpower to make the decision, or on the presence of a Frankfurt-style intervener who would prevent her from making the decision.[15]

5 Jeffrey's model

In [Jeffrey 1965: ch.11] and [Jeffrey 1968], Jeffrey observed that the problem of options resembles a problem in Bayesian models of learning or perception. The classical Bayesian account of learning assumes that the agent becomes absolutely certain of some evidence proposition E, comprising the totality of what the agent learns. But what is that proposition, in ordinary situations? When I look at the cloudy sky, I do not become certain that it is cloudy, irrespective of my background beliefs. (Recall that I give some credence to the hypothesis that I'm a brain in a vat.) Here, too, it is tempting to retreat from external-world propositions to propositions about a more secure internal world: perhaps what I learn with certainty is that it *appears to be cloudy*. On reflection, however, it is not clear that this helps. Couldn't we be mistaken or unsure about how things seem? Worse, ordinary propositions about how things seem are far too unspecific to explain the change in our beliefs prompted by a typical perceptual experience.

Jeffrey famously offered a different response. He suggested changing the classical model of learning to allow for cases in which the learning event may directly confer non-trivial probabilities $x_1, \ldots, x_n$ to a range of (mutually exclusive and jointly exhaustive) propositions $E_1, \ldots, E_n$, without rendering anything certain. Looking at the sky, for example, might make me 95% confident that it is cloudy. This kind of belief update has come to be known as *Jeffrey conditionalization* and is now a staple in Bayesian epistemology.

15 In [Hedden 2012: 352f.], Brian Hedden argues that construing options as decisions avoids the problem of internalism. (He does not consider the cover problem.) His argument relies on a tendentious conception of what it means that an agent *can decide* to ϕ. For example, he favourably considers the idea that an agent can decide to ϕ just in case she is not certain that if she decided to ϕ then she would fail to ϕ. This makes abilities to decide an internal matter. (Hedden claims that it even makes decisions-as-options satisfy Modal Certainty, assuming – without argument – that rational agents are always certain about their own credences.) But the proposed interpretation of 'can decide' is changing the topic. On this interpretation, it may well be that an agent "can" make a decision even though she is biologically incapable of making the decision. But arguably we don't want decision theory to predict or recommend choices the agent is biologically incapable of making.

Jeffrey made a parallel, though largely forgotten, proposal about an agent's options. Recall that on the classical model of deliberation, choosing an option goes along with becoming certain of that option. Jeffrey suggests that choosing may instead go along with assigning non-extreme probabilities $x_1, \ldots, x_n$ to a range of (mutually exclusive and jointly exhaustive) propositions $O_1, \ldots, O_n$. This probability assignment, rather than any particular proposition, then represents the chosen option.

The marksman's choice of a particular posture, for example, might make him 30% confident that *he will shoot and hit*, and 70% that *he will shoot and miss*. Other available postures would correspond to other probabilities. According to Jeffrey, these probabilities represent the marksman's options.

If we follow Jeffrey, the expected utility of an option can no longer be defined (as in section 2) in terms of the agent's credences conditional on the option, since options are no longer single propositions. We would need to generalize the usual definitions of expected utility.[16] This isn't trivial, but let's assume for the sake of the argument that it can be done. There is a more fundamental problem.

We want to know what options are available to an agent in a concrete choice situation. We're assuming that there are a number of acts the agent can perform, in the sense of section 2. The question is how these acts should be represented as decision-theoretic options. In Jeffrey's model, an option is a probability function over some partition of propositions. So our question is how the available acts map onto such probability functions.[17]

The mapping depends on the agent's background beliefs. For example, the marksman's credence that he will hit given his choice of a particular posture depends on his beliefs about the wind and the quality of his rifle. So we need to explain how the acts an agent can perform, in combination with her prior beliefs, determines the relevant "Jeffrey options": the available probability functions over partitions of propositions. And that turns out to be rather difficult.[18]

To see why, consider a simpler agent, of the kind studied in artificial intelligence: a robot with an internal representation of its environment, some goals, and a capacity to move around. If the robot's decision module figures out that moving to the left would be useful, a command is sent to its motor system that would normally cause the robot to move to the left. Whether the robot succeeds in moving to the left, of course, depends on various external factors such as the slipperiness of the floor or the presence of a glass wall

16 Perhaps the same generalization is required to handle "mixed strategies" in game theory.

17 To fully escape the cover problem, the mapping would have to be one-one. It is not obvious why that should be the case. However, one might argue that if two possible acts agree in their effects on the agent's beliefs, then they always have the same expected utility, so that it doesn't matter if we can't distinguish them.

18 The problem in some way resembles the "input problem" for Jeffrey conditionalization, discussed e.g. in [Field 1978], [Garber 1980], [Christensen 1992], and [Weisberg 2009].

blocking the way. Let's say the robot assigns probability 0.2 to such interfering factors. *Moving left* should then not count as one of its options (for the same reason for which *eating a blueberry* is not an option in the berry scenario from section 2). Following Jeffrey, the robot may instead have an option of *moving left* with probability 0.8 and *staying in place* with probability 0.2. But how is the robot's decision module supposed to figure out that this is one of its options—the option it could realize by sending a specific motor command? If the robot receives information increasing the probability of a slippery floor, the robot's options must change. Perhaps it now only has an option of *moving left* with probability 0.5 and *staying in place* with probability 0.5. Where do these probabilities come from?

They can't be simply be hard-coded into the robot's cognitive architecture. Suppose we programmed the robot so that the option that would be realized by sending the motor command for moving to the left is *moving left* with probability $Cr(\neg \mathit{Wall} \wedge \neg \mathit{Slippery}) + 0.5 \times Cr(\neg \mathit{Wall} \wedge \mathit{Slippery})$ and *staying in place* with probability $Cr(\mathit{Wall}) + 0.5 \times Cr(\neg \mathit{Wall} \wedge \mathit{Slippery})$. That would get the present example right, but it doesn't generalize. The mapping from acts to Jeffrey options shouldn't be fixed. Intuitively, the robot should be able to figure out under what conditions issuing the relevant motor command is likely or unlikely to have the desired effect, and it should be able to revise these judgements in light of its experience. How is that supposed to work?

6 Inventing options

I will now (finally!) present my own proposal for how to model decision-theoretic options. The proposal is a kind of hybrid between the options-as-decisions model from section 4 and Jeffrey's model from the previous section, but it adds an important twist.

Let's stay with artificial agents for a moment.

The task of a robot's decision module is to select a motor command appropriate to the robot's goals and information. In a sense, then, the options are simply given by the available motor commands. Suppose we use these motor command as the robot's options. That is, suppose each available motor command X is represented by an element $I(X)$ in the domain of the robot's subjective probability measure. (Motor commands are electrical signals, not propositions, so I assume they aren't themselves objects of probability.) If L is the motor command that would normally cause movements to the left, we could let the robot give high prior probability to *moving left* conditional on $I(L)$, but now the robot can learn under which conditions choosing $I(L)$ really leads to movements to the left and under which it doesn't. For example, the robot might observe that whenever there is a red light and it chooses $I(L)$, it doesn't end up moving to the left, so that it will learn to try other things when it sees the light.

In rough outline, this is indeed the standard approach in artificial intelligence (see e.g.

[Russell and Norvig 2010: ch.s 2 and 16]). Here, agents are assumed to have a "transition model" defining conditional probabilities over new states of the world given a present state and a given motor command. For example, if S is a slippery-floor state where our robot is at a certain place x, S' is a similar state where the robot is further to the left, and L is the motor command for moving to the left, the transition model might assign middling probability to S' conditional on S and L (or rather, $I(L)$, although the distinction between L and $I(L)$ is rarely made in artificial intelligence). These conditional probabilities can be adjusted through learning: given partial observation of a post-action state S', the robot can update its beliefs about the likely effect of a given motor command in a given pre-action state S.

I want to suggest that we should understand all rational decision-making along these lines. To be sure, human decision-making is more complex. For one thing, when we choose to move our limbs, a whole hierarchy of control systems appears to be in play. This presents an independent challenge to classical decision theory, which assumes a single interface between beliefs and desires on the one hand and actions on the other. The challenge could be met in different ways; I'll choose a lazy response of idealising away the hierarchy, treating it as a lower-level mechanism for selecting the final motor commands. We can also make decisions whose output isn't a motor command, as when we decide to travel to Paris, to visualize a scene, or to focus on our breath. To accommodate these decisions, we need to generalise the notion of a motor command. I'll simply speak of *outputs*.

The idea is that when we make a decision, our beliefs and desires combine to produce a specific neurophysiological event which in turn may cause body movements, changes of attention, new commitments, or whatever. The neurophysiological event is an "output" of the decision process.

The nature of the available outputs is of no concern to decision theory. We simply take it as given that there is a certain range of available outputs – of outputs that would be produced by some variation of the agent's volitional state. Each of these outputs X, I suggest, should be represented by an element $I(X)$ in the domain of the agent's probability function. These elements are the agent's options. If the agent finds that $I(X)$ maximizes expected utility, thus deciding to choose that option, the corresponding output X is produced.

When I say that $I(X)$ is a *proposition*, I mean that it is an element in the domain of the agent's credence function. Propositions, in this sense, are often construed either as sets of possible worlds or as sentences in a suitable language. On the sentential account, the option propositions should be *new* sentences, not definable in terms of ordinary sentences about the external or internal world. In particular, $I(X)$ should not be a detailed and accurate description of the corresponding output X.

Why not? Suppose in our own cognitive system the outputs that normally cause our arms to go up involve the release of glutamate. If the corresponding option propositions were detailed descriptions of these outputs in electrochemical terms, then on the classical model of deliberation choosing to raise your arm should make you certain that glutamate is being released in your brain. In reality, if you are ignorant or unsure about the role of glutamate in the neurochemistry of decision, merely deciding to raise your arms will hardly resolve your ignorance. I don't want to classify that as a failure of ideal Bayesian rationality.

Above I said that each output X should be *represented* by a proposition $I(X)$ in the domain of the agent's probability function, but that sense of 'representation' is merely causal: what's important is that if the agent chooses the option proposition $I(X)$, then the corresponding output X will be produced. The agent needs some way to pick out the relevant output, but she doesn't need to pick it out by its neurophysiological properties.

So if we had to design an ideal decision maker, we should expand the language over which the agent's credences are defined. We should add a new range of sentences for the specific purpose of representing options. In principle, it doesn't matter what these sentences look like; they could be atomic tags: 'X', 'Y', 'Z', etc.

Of course, if the agent has no idea which ordinary propositions are likely to be true if she chooses 'X' rather than 'Y', she will have no basis for choosing one over the other. (I assume the agent does not assign basic value to the option propositions themselves.) But even if the agent is initially ignorant of the connection between option propositions and ordinary propositions – a poor decision choice – she could easily acquire such information. For example, she could observe that choosing 'X' usually causes movements to the left. And she could learn how the effects of choosing 'X' depend on the inclination and slipperiness of the ground.

What is the meaning of the newly added sentences? It depends on what we mean by 'meaning'. There is a good sense in which the meaning of $I(X)$ is the corresponding output X, although meanings in this sense are not transparent to the agent. Alternatively, if choosing $I(X)$ usually leads to movements to the left, one might say that it represents *moving to the left*, in roughly the sense in which linguists sometimes take the imperative 'move to the left!' to mean that the addressee moves to the left. In yet another sense, $I(X)$ does not have a meaning at all. It doesn't need a meaning because its cognitive function is not to represent a certain kind of fact. Its cognitive function is to help the agent select a suitable output based on its internal model of the world. Again, what's important is that choosing $I(X)$ causes the agent's cognitive system to produce the corresponding output X. In principle, that's all you need to know about $I(X)$ to build or interpret a rational decision maker.

What if we understand the objects of credence not as sentences, but as sets of possible worlds? In that case, $I(X)$ can't be the set of worlds where output X is produced – unless we count ignorance of our neurochemistry as a failure of Bayesian rationality. And it can't be the set of worlds where the agent moves to the left, for the agent may not be certain that choosing $I(X)$ will lead to movements to the left. $I(X)$ can't be any set of ordinary, "metaphysically possible" worlds at all. More generally, options propositions can't be genuine *ways a world might be*. In many cases, any way a world might be will be either too strong, containing information the agent needn't have if she chooses the option, or too weak, failing to distinguish the option from all its alternatives.

So if you thought the objects of credence should be understood as ways a world might be, or as sets of metaphysically possible worlds, the problem of options calls for a revision. *Some* objects of credence may be genuine ways a world might be. Others are, well, options: further elements in the domain of an agent's credence function causally associated with the available decision outputs, as described above.[19]

You may balk at the idea that the objects of credence or belief include "propositions" that aren't ways the world could be. Fair enough. I don't mind if these things are called 'propositions', or if the expanded probability function is called a 'credence function'. We need the extra elements, and the expanded function, whatever they are called. If we restrict our attention to an agent's credences over ways the world might be (or ordinary sentences), we will only see what Jeffrey saw: different options will correspond to different shifts in an agent's credence function, for the most part without making any proposition certain. We won't be able to explain where these shifts come from and how they change through experience.

7 Options and actions

Let me explain how the model I have outlined is meant to be used, and how it deals with the problems we have encountered.

When we use decision theory to model a choice situation, we assume that there is a range of acts that are available to the agent, in something like the sense I tried to explicate in section 2. The problem of options is how these acts should be represented for the purpose of computing expected utilities. My answer is simple: each available act is represented by a primitive option proposition. Functionally, option propositions correspond to action-initiating outputs of the agent's decision process: if a rational agent finds that a given option proposition maximizes expected utility, the corresponding output (a motor command, for example) is produced. But this correspondence is merely causal. Option propositions do not represent the relevant acts in terms of their lower-level physiological beginnings. They don't have any direct, categorical content at all; they aren't ways a world might be, but they stand in probabilistic relations to different ways

19 The objects of credence are standardly assumed to form a sigma-algebra. Starting with an ordinary sigma-algebra A, the expanded algebra is the algebraic product of A and the algebra generated by the atoms $\{I(X) : X \text{ is an output}\}$.

a world might be. (If you insist on assigning them content, their content will be a probability measure over such ways – a measure that changes with changes to the agent's background beliefs.)

The possible "outputs" of deliberation are what I called "motivational variations" in section 2. I can't give you a simple, straightforward procedure for determining the relevant outputs for a concrete agent in a concrete decision context. Nor can give you such a procedure for determining the agent's credences and utilities. I want to treat these questions alike, as questions of modelling. There are no hard and fast rules for applying abstract, high-level models to concrete situations in the world.

Here is how the present model could be applied to the berry scenario from section 2. The choice you face is, in fact, a choice between eating a blueberry and not eating a blueberry. But that's not how the choice is presented to you, since you aren't sure if the berry is a blueberry; you may not even be sure if there is an external world. From your perspective, the option you choose is a primitive option proposition of which you are, say, 95% certain that it will lead to eating a blueberry, with most of the remaining credence going to the hypothesis that it will lead to eating a tutsan berry.[20] Other options are, according to your beliefs, practically certain to make you not eat any berry. The decision matrix for the two kinds of options may look as follows – O_1 is an option that (in fact) leads to eating a blueberry, O_2 is an alternative that doesn't.

	Blueberry	*Tutsan Berry*	*No External World*
O_1	1	-4	-1
O_2	0	0	-1

O_1 maximizes expected utility as long as your credence in the *Blueberry* hypothesis is greater than 80%.

Notice that there is no need to specify the physiological "output" corresponding to the different options. The information I have given is enough to predict (or recommend) that you will eat the blueberry. We can also see why, in practice, it does little harm to represent the two options as *eating the berry* and *not eating the berry*: since both options have equal utility in the *No-External-World* state, that state can be ignored for the purpose of comparing expected utilities. Similarly, even though you could realize *eating the berry* and *not eating the berry* in many different ways – using your left hand or your right hand, etc. – within reason these differences plausibly make little difference to expected utility; so there's no need to look at all the fine-grained options you actually have.

20 I don't mean that you consciously conceive of the chosen option as a primitive option proposition. Remember that decision theory is not a theory of conscious deliberation.

The option propositions in my model go well with the classical model of deliberation. Since they aren't genuine ways a world might be, we don't need to worry how agents could become absolutely certain of such propositions merely through deliberation. We can accept the classical model of deliberation without assuming that deliberation provides infallible access to special facts about the world.

My option propositions are also tailored to meet the Cover condition. If the marksman can choose between thousands of torque configurations, he will be modelled as having thousands of option propositions in his extended doxastic space. When we were looking for option propositions among genuine ways the world could be, we were stuck between the competing demands of, on the one hand, making the propositions so weak that they don't entail external facts of which the agent is unsure, while, on the other hand, making them so strong that they exclude one another. The introduction of quasi-propositions as options makes it easy to satisfy both demands. Option propositions are disjoint without entailing any external facts about the world.

What about the problem of internalism? Won't the available option propositions depend on cognitively external facts – facts about which "outputs" are available? To some extent, yes. Each option proposition should correspond to an output the agent is, in principle, physiologically capable of producing. There shouldn't be "dangling" option propositions not associated with any possible decision output for the relevant kind of agent. But that is simply a constraint of the model. If an agent is adequately designed in accordance with the model, or represented in terms of the model, there won't be any dangling options.

There may also be options the agent is incapable of choosing due to fleeting external factors such as the presence of a Frankfurt-style intervener. But in cases like that, we arguably *should* include the relevant option, and instead restrict the assumption that rational agents always choose options that maximize expected utility. Concretely, suppose a Frankfurt-style intervener would prevent you from choosing O_1 in the berry situation. You're an ideal Bayesian agent, but have no idea of the demon's presence. If O_1 would maximize expected utility, decision theory arguably shouldn't predict that you will choose an alternative option. Rather, it should predict that you'll be stopped by the demon.[21]

To conclude, let me briefly return to Jeffrey's insight that our choices are part of the

21 Things get worse if you are not oblivious of the demon. Suppose you suspect a demon will strike you down if she detects that you're about to choose O_1. Intuitively, you now have a reason against O_1. But how should we model this? The problem is that choosing O_1 is incompatible with the presence of the demon (because the demon would have prevented O_1): $Cr(\mathit{Strike}/O_1) = Cr(\mathit{Strike}\backslash O_1) = 0$. But then how does the demon provide a reason against O_1? (Similar problems arise in cases where mixed decision equilibria are punished.) Introducing "virtual" option propositions helps make decision theory applicable to a wide range of ordinary cases, but it doesn't make it applicable to every conceivable case.

natural world and should therefore be represented as ordinary propositions. The model I have proposed suggests that this is not quite right. From the decision-maker's perspective, an object of choice is a proposition outside the natural realm, only contingently and probabilistically related to propositions about the natural world. This might be taken to explain or even vindicate the widespread intuition that rational decision-makers must treat their choices as 'interventions', not governed by ordinary physical laws, and not constrained by degrees of belief pertaining to natural propositions. Whether the model I have outlined really does vindicate these intuitions is a question I have to leave for another occasion.

References

G.E.M. Anscombe [1957]: *Intention.* Ithaca, NY: Harvard University Press

Michael Bratman [1987]: *Intention, plans, and practical reason.* Cambridge, MA: Harvard University Press

Campbell Brown [2018]: "Maximalism and the Structure of Acts". *Noûs.* Forthcoming

David Christensen [1992]: "Confirmational Holism and Bayesian Epistemology". *Philosophy of Science*, 59(4): 540–557

Hartry Field [1978]: "A Note on Jeffrey Conditionalization". *Philosophy of Science*, 45(3): 361–367

Harry Frankfurt [1969]: "Alternate possibilities and moral responsibility". *The Journal of Philosophy*, 66(23): 829–839

Daniel Garber [1980]: "Field and Jeffrey Conditionalization". *Philosophy of Science*, 47(1): 142–145

Allan Gibbard and William Harper [1978]: "Counterfactuals and Two Kinds of Expected Utility". In C.A. Hooker, J.J. Leach and E.F. McClennen (Eds.) *Foundations and Applications of Decision Theory,* Dordrecht: D. Reidel, 125–162

Brian Hedden [2012]: "Options and the subjective ought". *Philosophical Studies*, 158(2): 343–360

Frank Jackson [1991]: "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection". *Ethics*, 101(3): 461–482

Richard Jeffrey [1965]: *The Logic of Decision.* New York: McGraw-Hill

— [1968]: "Probable knowledge". *Studies in Logic and the Foundations of Mathematics*, 51: 166–190. Reprinted with minor revisions in [Jeffrey 1992]

— [1992]: *Probability and the Art of Judgment.* Cambridge: Cambridge University Press

James Joyce [1999]: *The Foundations of Causal Decision Theory.* Cambridge: Cambridge University Press

— [2012]: “Regret and instability in causal decision theory”. *Synthese*, 187(1): 123–145

Gregory S. Kavka [1983]: “The toxin puzzle”. *Analysis*, 43: 33–36

David Lewis [1979a]: “Attitudes *De Dicto* and *De Se*”. *The Philosophical Review*, 88: 513–543

— [1979b]: “Counterfactual Dependence and Time’s Arrow”. *Noûs*, 13: 455–476

— [1981]: “Causal Decision Theory”. *Australasian Journal of Philosophy*, 59: 5–30

Patrick Maher [1993]: *Betting on theories*. Cambridge University Press

John Perry [1979]: “The problem of the essential indexical”. *Noûs*, 13: 3–21

Philip Pettit [1991]: “Decision Theory and Folk Psychology”. In M. Bacharach and S. Hurely (Eds.) *Foundations of Decision Theory: Issues and Advances,* Cambridge (MA): Blackwell, 147–175

John L. Pollock [2002]: “Rational Choice and Action Omnipotence”. *The Philosophical Review*, 111: 1–23

Stuart J. Russell and Peter Norvig [2010]: *Artificial Intelligence: A Modern Approach*. Cambridge (MA): MIT Press, 3rd edition

Leonard Savage [1954]: *The Foundations of Statistics*. New York. Wiley

Wolfgang Schwarz [2018]: “Imaginary Foundations”. *Ergo*

Brian Skyrms [1980]: “Higher Order Degrees of Belief”. In D.H. Mellor (Ed.) *Prospects for Pragmatism,* Cambridge: Cambridge University Press

— [1984]: *Pragmatics and Empiricism*. Yale: Yale University Press

— [1990]: *The Dynamics of Rational Deliberation*. Cambridge (Mass.): Harvard University Press

Jordan Howard Sobel [1971]: “Value, Alternatives, and Utilitarianism”. *Noûs*, 5(4): 373–384

— [1980]: *Probability, Chance and Choice*. Unpublished book manuscript

— [1983]: “Expected utilities and rational actions and choices”. *Theoria*, 49: 159–183. Reprinted with revisions in [Sobel 1994: 197–217]

— [1986]: “Notes on decision theory: Old wine in new bottles”. *Australasian Journal of Philosophy*, 64: 407–437. Reprinted with revisions in [Sobel 1994: 141–173]

— [1988]: “Infallible Predictors”. *The Philosophical Review*, 97(1): 3–24

— [1994]: *Taking Chances*. Cambridge: Cambridge University Press

Jean-Paul Vessel [2003]: “Counterfactuals for consequentialists”. *Philosophical Studies*, 112(2): 103–125

Paul Weirich [1983]: "A decision maker's options". *Philosophical Studies*, 44(2): 175–186

Jonathan Weisberg [2009]: "Commutativity or holism? A dilemma for conditionalizers". *The British Journal for the Philosophy of Science*, 60(4): 793–812