Options and Actions

Wolfgang Schwarz*
Draft, 17 September 2014

1 Introduction

Decision theory says that rational agents should choose acts that maximize expected utility with respect to their subjective probability and utility functions. But even if we know the probabilities and utilities of an agent in a given decision situation, this does not allow us to predict what the agent will do, unless we also know what acts are available. What are the options from which she can choose?

The question is especially pressing if we understand decision theory, as I will throughout the present paper, as a psychological model – as part of an idealized, high-level, computational design. Such a model may, in turn, be understood in different ways. It could be a descriptive model, trying to capture, to a first approximation, the choices made by actual humans. Or it could be a normative model, specifying the choices people should ideally 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 acts are simply given as part of a well-defined decision problem. Somehow the agent's cognitive system must itself figure out the available options.

This leads to problems. For example, we can hardly assume that the agent has infallible access to what acts she can perform. Suppose she believes her alternatives are A and B, while in fact they are B and C. Suppose also that each of A and C has greater expected utility than B. We don't want to predict that the agent does C, if her decision mechanism isn't aware that C is an option. But arguably we also don't want to predict that the agent does A, which is guaranteed to be false, since the agent can't in fact do A.

Reflecting on this and other problems, I will suggest that we need an intermediate layer between beliefs and desires on the one hand and acts on the other. The output of a decision process selects an element of this intermediate layer; for the purpose of computing expected utilities, these elements are the agent's options. They are not acts the agent can perform, although they are probabilistically related to such acts. In the agent's cognitive system, they are best represented as primitive propositions.

* Acknowledgements

2 What we can do

Let's say that an agent's *objective options* are the acts she can perform in a given decision situation. I will argue that this conception of options is not adequate for the purpose of computing subjective expected utilities. On the other hand, at some point decision-theoretic models of choice obviously have to make contact with the agent's objective options, so it may be worth thinking a bit about options in the objective sense. (I will return to this topic at the end of the paper.)

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 theoretic algorithms are 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. Not so when the robot reaches a junction. Here the outcome is under control of the robot's internal state: vary the state, and the robot will choose different paths. That is all the freedom we need.

To make room for that much freedom, we must not hold fixed an agent's total psychological state when we determine her objective options. Given your actual beliefs and desires and the way you reach decisions, it may be impossible that you turn right. But this doesn't mean that turning right is not one of your options. As a vegetarian, I might say that I *can't* eat the mince pie I am offered, or that eating it is not an option for me. But again this is not the relevant sense for the purpose of decision theory. The reason why I won't eat the pie lies in my beliefs and preferences, not in the fact that the act is not available to me in the first place.

As a first stab towards analysing objective options, we might say that A is an (objective) option for an agent in a given situation iff some variation of the agent's psychological state would make her perform A. But this is a little too wide. It would falsely predict that one of my present options is to believe that Sydney is the capital of Australia. After all, there is a variation of my psychological state – notably my state of belief – that would make it true that I believe that Sydney is the capital of Australia. But arguably my beliefs are not in the right sense under my psychological control. We need to focus on more specific variations of the agent's psychological state.

The classical counterfactual analysis of 'can' gets things roughly right. On this account, an agent *can* A just in case she *would* A if she *intended* to A. Accordingly, we could define objective options (for an agent at a time) as things the agent would do if she intended (at the time) to do them. I prefer a slightly weaker alternative that retains the existential form of our first stab: A is an objective option if some variation of the agent's intentional state (at the time) would bring about A.

Neither of these is adequate as a general analysis of the English auxiliary 'can'. For example, they don't capture the sense in which I *can't* eat the mince pie. Nor do they capture the sense in which I *can* play the piano even when there is no piano around (in which case it is not true that if I intended to play the piano, I would play). However, for our topic these verdicts are exactly right. If there is no piano around, then playing the piano is not one of my options.

What about another stock example against the counterfactual analysis: a person in a coma who can't do anything, although we may assume that she would get up if she intended to get up (because if she intended to get up she would no longer be in a coma)? If getting up is one of her options, won't decision theory falsely predict that the patient ought to get up (assuming she still has beliefs and desires)? Perhaps it will, but clearly decision theory is not a complete psychological model for agents who can fall into a coma. Arguably, one of the things that happen in a coma is precisely that the agent's decision module no longer works: she no longer chooses options with greatest expected utility. So it doesn't really matter what options we attribute to her. More generally, decision-theoretic models describe the workings of an idealized cognitive system. They don't apply to rocks or corpses or people in a coma. Decision theory is not meant to predict or evaluate the behaviour of these systems, so we may well set aside the question how to delineate their options.[1]

These observations touch on a possible disconnection between decision-theoretic options and everyday normative attitudes towards people's actions. One might have thought that when we ask what someone ought to do – from a moral perspective, perhaps, or to achieve certain goals –, then the relevant alternatives are precisely her decision-theoretic options. But I don't think that's generally true.

For one thing, when we give advise or attribute blame and praise, we often hold fixed various facts that should not be held fixed when we catalogue the space of decision-theoretic options. Thus we exclude options that would go against deeply ingrained fears or hopes or convictions: since you can't bring yourself to be friendly, you ought to be

1 To block other, more problematic, counterexamples, we might have to stipulate that the counterfactuals in our analysis must not be read in a "backtracking" manner (see [Lewis 1979]). Imagine an agent who has no reason to raise her hand and narrowly escapes a stroke-like event that would have (a) made her want to raise her hand, and (b) paralysed her arms. One might intuit that *if the agent had intended to raise her hand, she would have failed to raise her hand*, despite the fact that raising her hand should count as one of her options. But here the italicised counterfactual is backtracking.

polite. Perhaps we can even hold fixed quite commonplace intentions: since you intend to change lanes, you ought to accelerate (see [Goldman 1978]). There is also a temptation to hold fixed the past and the laws of nature, which is why some people intuit that determinism renders praise and blame pointless: you couldn't really have done anything else.

In these respects, the range of options that figure in everyday normative evaluations is often more narrow than decision-theoretic options. In other respects, it can be wider. In particular, decision theory as a psychological model should not consider entire strategies, i.e. sequences of actions, unless the agent has some way of "binding" herself to the sequence. To illustrate, consider Professor Procrastinate (from [Jackson and Pargetter 1986]), who is asked to write a review, but probably won't actually write it if he agrees. Assume the professor has no means of binding: there is nothing he can do, right now, at the point when he is asked, that would ensure that his future self will write the review. *Agreeing and then writing the review* is then not one of his decision-theoretic options, and it doesn't pass our counterfactual analysis.[2] From a normative perspective, the situation is more complex. Normative evaluations often pertain not only to individual choices, but also to sequences of choices. If the professor's intentional state were different now as well as in the future, he would write and complete the review. In this sense it is in his power to accept and write, and one might reasonably hold that this is what he should do.

The upshot is that we should not – as many authors do – equate the notion of options relevant to decision theory (as a psychological model) with the notion of options relevant to everyday normative evaluation. Again, there is surely a connection between the two topics. One might suggest that there is a special ("deliberative", "subjective", "ideal") sense of *ought* for which the relevant options are precisely the options of decision theory. I have sympathies for this idea. But before we consider it any further, we should get clear on what counts as an option in decision theory, without presupposing a particular connection to our normative practice.

So far, we have considered a decision-makers options "from the outside", as the acts she can in fact, objectively, perform – on a specific reading of 'can'. One might think that an agent who follows a decision-theoretic model of choice will simply choose whichever of these acts has greatest expected utility. The most obvious problem with this idea is that it doesn't take into account the agent's own information about what acts she can perform. Before I turn to that, I want to address another problem: that ordinary acts aren't the kinds of things whose expected utility is even well-defined.

2 Here I side with Joyce [1999: 57–61] against Savage [1954: 15f.]. However, Savage may not think of decision theory as a psychological model in the present sense, so the disagreement may be superficial.

3 Decisions and propositions

Decision theory requires us to trace the same options across a range of hypothetical circumstances. Imagine you are preparing an omelette and consider adding another egg, of which you're not sure whether it is still good. Let's say you in fact go ahead and add the egg, which is rotten. So the act you perform, call it α, is an act of adding a rotten egg to the omelette. To compute the expected utility of your choice, we have to consider its outcome in all relevant states of the world. Thus if α is the option you chose, then we'd have to consider the outcome of performing α in situations in which the egg is not rotten. But could you have performed α – that act you actually performed – if the egg had not been rotten? One might intuit that the answer is no: if the egg had been good, you would have performed a different act by adding it to an omelette, and act that doesn't amount to spoiling the omelette.

To avoid problems like this, decision-theoretists generally don't identify options with acts in the ordinary sense. Savage [1954] construes options ("acts") as abstract functions from possible states of nature to outcomes. Jeffrey [1965] put forward a more elegant proposal on which options are simply propositions of a certain kind. Your token act of adding an egg may be identical to your act of adding a rotten egg, but the proposition *that you add an egg* is undoubtedly different from the proposition *that you add a rotten egg*. Only the former provides a reasonable characterisation of the option you chose.[3]

In contrast to Savage, Jeffrey drops the sharp distinction 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]. His proposal opened up the possibility of considering probabilities of options, probabilities conditional on options, or probabilities of complex propositions involving options, which is put to use not only in Jeffrey's own formulation of decision theory but also in its main rivals (such as [Gibbard and Harper 1978], [Lewis 1981], [Sobel 1986], [Skyrms 1984] and [Joyce 1999]).[4]

A further advantage of identifying options with propositions is that we can then allow for options that may not be adequately described as acts at all. In Kavka's [1983] toxin puzzle, you arguably don't just face a decision between *drinking the toxin* and *not drinking the toxin*, but also between *intending to drink the toxin* and *not intending to drink it*. One might be reluctant to call an episode of intending to drink the toxin an *act*, but the existence of a corresponding proposition should be unproblematic. (In section 6, we will consider the view that options are always intention-describing propositions rather than act-describing propositions.)

3 More reasons for individuating acts as propositions can be found in [Bennett 1988] and [Bennett 1995].

4 While most decision theorists in philosophy have adopted Jeffrey's construction of options, some hold that deliberation is incompatible with having credences about one's options. We will in fact come to reconsider Jeffrey's insight. On a certain understanding of 'credence', the model I will propose even agrees that options are not suitable objects of credence.

A minor verbal complication with the construction of options as propositions is that propositions are not the kinds of things one can *do*. I will use the construction *S makes true A* to express that an agent realizes an option. The precise semantics of that phrase won't be important; indeed, it will do no harm to read *S makes true A* as equivalent to *A*: you make true that you turn left iff you turn left.

If we adjust the analysis of objective options from the previous section to the propositional framework, we get something like Lewis's definition of options in "Causal Decision Theory" [Lewis 1981]:

> Suppose we have a partition of propositions that distinguish worlds where the agent acts differently [...]. Further, he can act at will so as to make any one of these propositions hold, but he cannot act at will so as to make any proposition hold that implies but is not implied by (is properly included in) a proposition in the partition. [...] Then this is the partition of the agent's alternative *options*. [Lewis 1981: 308]

Lewis here identifies the agent's options with act-describing propositions the agent can make true ("at will"). Note that if an agent can make true a proposition *A* – say, on the counterfactual readings of 'can' from the previous section – then she can also make true any proposition entailed by *A*. If you can make it true that you raise your arm, then you can also make it true that you move a limb. Lewis's definition excludes such unspecific options. To see why this might be useful, consider a situation in which you are punished for intentionally raising your arm but rewarded for unintentionally raising it. Suppose you have no means of causing yourself to unintentionally raise your arm, and so you (rationally) don't raise your arm. In this case, it might well be that the most likely circumstances under which you'd raise your arm are circumstances in which you do so unintentionally. As a consequence, *raising your arm* then has high expected utility. If we count it as one of your options, we might wrongly conclude that you should choose it. However, in fact your only relevant option is the more specific *intentionally raising your arm*, which has low expected utility.[5]

5 Lewis does not motivate his definition at all. Explicit discussions of the specificity issue (as in [Weirich 1983] and [Sobel 1983]) mostly focus on cases where an unspecific option is an (exclusive) disjunction of more specific options. This makes the issue very subtle. For example, it may seem that a disjunctive option has maximal expected utility only if all its disjuncts do, in which case including the unspecific option would at least do no harm. This is certainly true in Jeffrey's "evidential" decision theory. On the other hand, Sobel [1983: 203–208] presents a complicated scenario in which a disjunctive option seems to have greater "causal" expected utility than all its disjuncts.

Another reason for excluding unspecific options in causal decision theory is that the rules for evaluating causal expected utility become problematic if the relevant options are very unspecific: subjunctive conditionals and conditional chances are often hard to evaluate if the "antecedent" proposition is unspecific.

Be that as it may, the real defect with Lewis's proposal is that it still defines an agent's options "from the outside", which won't do if we want to interpret decision theory as a psychological model.

To illustrate the problem, suppose you find yourself in a hotel room and want to turn on the lights. You notice a button at the wall which you suspect functions as a light switch – as in fact it does. Assuming that you "can act at will so as to make hold" the proposition *that you push the button*, presumably you can also act at will so as to make true the stronger proposition *that you turn on the light by pushing the button*. (It certainly seems that you can make true that proposition on both of the counterfactual analyses considered in the previous section.) By Lewis's definition, pushing the button is then not one of your options, since it is entailed by the more specific option of turning on the light by pushing the button. In any case, the more specific proposition counts as an option. But given your state of uncertainty, it shouldn't. Suppose you give some probability to the hypothesis that the button's function is to call room service, which you don't want to do. When you evaluate your options, you should then consider the consequences of pushing the button both on the assumption that it is a light switch and on the assumption that it calls room service. If you had the option to turn on the light by pushing the button, your decision problem would be much simpler: you would obviously choose that option.

Intuitively, the problem is that Lewis's definition makes an agent's options depend on external facts of which the agent may not be aware. While it is true that you can turn on the light by pushing the button, this is not something you know. Let us try to get a clearer grip on what is needed.

4 Access conditions

In response to cases like the lightswitch scenario, a natural thought is to impose the following condition on *subjective options* – i.e., the options whose expected utility determines what the agent should do (see e.g. [Hedden 2012: 348]).

(C1) A proposition is an option for an agent at a time only if the agent is certain that she can make it true.

Here 'can' must be read in the stronger, non-existential sense: the agent must be certain that the proposition would be true if she intended it to be true at the relevant time.[6]

6 To see why the existential sense won't do, suppose there are two buttons, and you know that one of them operates the lights while the other calls room service. You don't know which does which. Then you know that in the existential sense you can make true the proposition that you turn on the light by pressing a button. Yet in an adequate representation of your decision problem, none of your options should entail that proposition.

From (C1), it is a short step to a popular definition of subjective options on which an agent's options are the maximally specific propositions of which she is certain that she can make them true.

One problem with this line of thought is that it goes far beyond what cases like the light switch example support. We can grant that the range of options should be accessible to the decision-making process. But whatever exactly that means, it arguably doesn't require anything as strong as (C1). Compare the agent's degrees of belief and desire. These are presumably accessible, in the relevant sense, to her decision-making mechanism. They are so even if the agent is not certain about her own beliefs and desires.[7] Why must the agent be certain about what she can make true?

To be sure, some authors portray decision theory as the insane doctrine that rational decision-makers should consciously compute the expected utilities of their options. This would seem to require that the agent knows about her options, as well as her probabilities and utilities. But the proposed interpretation of decision theory makes the theory utterly implausible as a descriptive, normative or constitutive model. Real people don't explicitly compute expected utilities whenever they face a choice, nor is there any reason to think that they should. A decision-theoretic agent in the sense I'm interested in is an agent who reliably chooses options with highest expected utility (provided there are such options). Decision theory does not fix an algorithm for realizing that goal, and it certainly doesn't say that the algorithm must involve conscious operations of multiplication and addition.

So we should be cautious with assumptions like (C1). Indeed, it seems plausible to me that simply decision-theoretic agents need not have any opinions at all about what they can make true – that is, about what would be the case if they intended it to be the case. A decision-theoretic agent doesn't need to have the "concept" of an intention.

A further drawback of defining options as propositions of which the agent believes that she can make them true is that we would run into the issue I mentioned in the introduction: we would allow for options the agent actually *can't* make true. If you are falsely convinced that you can fly, and we only consider your beliefs about what you can bring about, decision theory might predict that you will (or should) fly. To avoid this, Sobel ([1980], [1986]) stipulates that the decision-maker is never mistaken when she is certain about something, but that is hardly a satisfactory answer if we're interested in a psychological model.

In any case, the most serious problem with (C1) is that many actual decision problems don't seem to provide suitable propositions that pass the condition. For example, at the moment I am not absolutely certain that my muscles will keep functioning during the next few seconds. Consequently, I am not absolutely certain that I can raise my hand in

7 Some decision theorists have argued that higher-order degrees of belief concerning one's own degrees of belief are incoherent. That would be sufficient to make my point. However, I agree with Skyrms [1980] that higher-order degrees of belief are unproblematic, and do not only take the values 0 and 1.

the sense that if I now intended to raise my hand, then I would end up raising it. Is there any non-trivial proposition of which I, *qua* decision-maker, must be absolutely certain that I can make it true – anything for which I must be absolutely certain that it would be true if only I intended it to be true? Arguably not.

In [Jeffrey 1968], Jeffrey (who elsewhere seems to endorse the above definition of options) discusses another condition on subjective options:

(C2) A proposition is an option for an agent in a decision situation only if it would be rational for the agent to become certain of the proposition merely by making a choice.

Like (C1), this excludes *turning on the lights by pushing the button* as an option in the light switch example, since it would not be rational for you to become certain that the lights will go on merely by deciding to push the button. In contrast to (C1), what is supposed to become certain in (C2) is only the actually chosen proposition, not any claim about the agent's ability to make true that proposition.

(C2) is better justified than (C1). What motivates it is the observation that decision-making provides "knowledge without observation", as Anscombe said. Having decided to order coffee, I won't be surprised to find myself placing the order. Merely by reaching the decision, I already knew that (unless unexpected circumstances intervene) I would go on and place the order. Anscombe's observation is supported by the models of deliberation outlined in [Skyrms 1990], on which rational deliberation systematically goes along with becoming certain that the eventually chosen proposition is true.

Unfortunately, (C2) is fairly useless as a criterion a decision process might use to identify an agent's options. How is the agent's cognitive system supposed to determine what it would be rational for the agent to believe merely by making a choice?

(C2) also appears to share the main problem with (C1): ordinary decision situations don't come with sufficiently specific propositions that satisfy the condition. Deciding to raise my arm would not make me absolutely certain that I will raise my arm, nor should it. So what are my options?

5 Uncertain options?

We seem to have reached an impasse. On the one hand, we need to take into account the agent's information when characterizing her options. On the other hand, it is hard to see how this can be done without requiring that the agent is in some sense absolutely certain about her options, which seems implausible.

Can we relax the certainty conditions? What if instead of (C1) we say that for a proposition to be an option, it is enough that the agent is *reasonably confident* that

she can make it true? Or, building on (C2), that it would be rational of her to become *reasonably confident* that the proposition is true merely by making a choice?

Well, return to the light switch scenario. Suppose you are almost certain that the button is a light switch, but reserve about 1 percent (or 0.1 percent) of your credence to the hypothesis that the button has been rewired to detonate a bomb. It would clearly be a mistake to say that one of your options is to turn on the lights by pushing the button – even though you are almost certain that this is something you can do, and even though your choice might rationally make you almost certain that it is true.

Of course, when we think and talk about situations in which an agent faces a decision, we often ignore remote possibilities, even if the agent's evidence does not rule them out. We take for granted that the agent's muscles will keep functioning, that she won't be abducted by aliens, that light switches are not wired to bombs, etc. But an ideal decision process should never ignore possibilities with positive probability.

Worse, often it is even hard to find suitable propositions of which the agent is reasonable confident. Jeffrey [1968: 37] considers the case of a marksman aiming at a distant target. What are the marksman's options? *Shooting* is too unspecific. The marksman can shoot in many different ways, varying the direction in which his gun is pointing, the height at which he holds it, etc., and these variations make a great difference to his probability of hitting the target. *Shooting and hitting*, on the other hand, is too specific. We can assume that there is indeed a particular way of shooting that would make it true that the marksman hits. But the marksman doesn't know which way it is. He isn't irrational if he misses. *Shooting in direction x at height y etc.* also won't do, for this is neither something of which the marksman believes that he can make true in the relevant sense (he doesn't believe that if only he intended to make it true, it would become true), nor is it something of which he can rationally become certain merely by making a choice.

Jeffrey's solution (also presented in section 11.9 of [Jeffrey 1965]) mirrors his account of perception. The idea is to model options not as single propositions, but as probability measures over a certain partition of logical space. In the example of the marksman, Jeffrey suggests that the chosen option might be given by the partition $\{$*shooting and hitting*, *shooting and missing*$\}$ with associated probabilities 0.3 and 0.7. Instead of making him certain of anything – as assumed by (C2) –, the marksman's choice will make him 30 percent confident that he will hit the target, and 70 percent confident that he will miss. We can also accommodate more remote possibilities in this manner. If the marksman gives credence 0.001 to the hypothesis that his gun will explode, we can use the partition { *shooting and hitting*, *shooting and missing*, *getting killed in the gun's explosion* }, with associated probabilities 0.3, 0.699, and 0.001 (say).

In Jeffrey's model, choosing an option is like choosing a lottery that will lead to the specified outcomes with the specified probabilities. The trick is that the lottery is not itself represented as a proposition. Curiously, Jeffrey therefore retracts his fruitful

construction of options as propositions in favour of an account on which options look a lot like the "prospects" or "lotteries" in earlier formulations of decision theory – except that the "pure" options over which the lotteries are defined are partitions of ordinary propositions.

On Jeffrey's account, (C2) can be weakened to (C2′).

(C2′) An assignment of probabilities $x_1, x_2, \ldots$ to the members of some (mutually exclusive an jointly exhaustive) propositions $A_1, A_2, \ldots$ is an option for an agent in a decision situation only if it would be rational for her to have credences $x_1, x_2, \ldots$ in $A_1, A_2, \ldots$ respectively merely by making a choice.

(C2′) no longer requires the agent to be, or become, certain of her options. However, for our purpose it can hardly count as a satisfactory definition of options. For one thing, (C2′) only gives a necessary condition, and it is not obvious how to turn it into a necessary and sufficient condition. More importantly, like (C2), (C2′) offers little guidance for a decision process to determine the agent's options. It may well be right that one of the marksman's options would rationally make him 30 percent confident that he will hit the target, and none of his options would rationally make him 50 percent confident. But this is clearly not a brute fact. As Jeffrey points out, "[t]he basis for his belief may be his impressions of wind conditions, quality of the rifle, etc." [Jeffrey 1968: 37]. If the wind increases, the marksman will no longer have an option that corresponds to probabilities 0.3 and 0.7 over $\{$*shooting and hitting*, *shooting and missing*$\}$. In general, the lotteries that represent available options for an agent seem to be determined by the agent's beliefs, but Jeffrey's account falls silent on how this works.

A parallel question arises in Jeffrey's account of perception. There, the problem is known as the "input problem" for Jeffrey conditioning, and widely thought to be unsolvable (see e.g. [Field 1978], [Garber 1980], [Christensen 1992], [Weisberg 2009]). My own view is that in either case, the problem is indeed unsolvable within the strict confines of Jeffrey's "radical probabilism". It can only be solved by reverting to the traditional picture on which there are propositions to which perception and action confer absolute certainty.

6 Intentions

Let's set aside Jeffrey's proposal for the moment and return to the idea that options are propositions, rather than lotteries over propositions. Can we find propositions that pass the certainty requirements introduced above? Propositions about overt acts typically won't do – there is always a small chance that our muscles won't work in the normal manner. But maybe we can retreat to more immediate and secure consequences of decisions: episodes of *trying*, *intending*, *willing*, or *deciding*. Ideas along this line have

been defended by several authors, including Weirich [1983], Sobel [1971], [1983], Joyce [1999], and Hedden [2012]. There are in fact good reasons to allow for intentions or decisions as options, quite independent of the present considerations.

Suppose you are in Damascus at noon and consider going to Aleppo in the evening. If *going to Aleppo in the evening* is one of your options, we should then ask what is likely to happen if you go to Aleppo. But suppose you're inclined to stay in Damascus, so that the most likely scenario in which you go to Aleppo is one in which you get drunk in the afternoon and then decide to go on a whim, without packing enough food and clothes and without informing your family. This scenario, we may assume, has very low subjective utility. Does that give you a reason against deciding (now) to go to Aleppo? Clearly not. If you now decide to go, you will of course make the necessary arrangements. Thus what matters are not the probable scenarios in which you *go to Aleppo in the evening*, but the probable scenarios in which you *now decide to go to Aleppo in the evening*. For the purposes of calculating expected utilities, this is the relevant option to consider.

Indeed, as I already mentioned above, sometimes the main reasons for deciding to perform an act are not consequences of the act itself, but only consequences of the decision. An anxious person may reasonably make a decision just to stop worrying and calm her mind, even if other considerations would support delaying the decision until more information becomes available. In the toxin puzzle, the reason for deciding to drink the toxin is that the resulting state of decision or intention will be rewarded. Similarly in the more realistic scenarios on which the puzzle is modelled. A firm decision to leave your partner if she betrays your trust can have high expected utility even though acting on the decision does not.

What is highlighted by these cases is that agents face decisions not only when they physically reach a junction in a road. We can make decisions not only concerning our immediate behaviour but also concerning our behaviour in the distant future. This is useful for a variety of reasons. In particular, it can provide valuable information about the future and thus allow for more informed decisions. Whether or not you will go to Aleppo in the evening makes a big difference to how you should spend the afternoon. Conversely, what you do in the afternoon affects the expected utility of then leaving to Aleppo. Deciding ahead of the time to go to Aleppo simplifies the problem and allows for more optimized courses of action, by reducing the uncertainty about your future behaviour.[8]

To allow for planning, it would not be enough to move from overt acts to tryings.

8 There are other advantages. The capacity of forming intentions, which goes hand in hand with a reluctancy towards overturning intentions, might provide Professor Procrastinate with a means of binding himself to a sequence of actions, which allows him to lead a life of greater subjective utility; the capacity is also useful in deterrence cases and the decision puzzles of [Arntzenius et al. 2004], and it might enable useful social institutions and conventions ([Lewis 1969]).

The most likely scenarios in which you *try to go to Aleppo in the evening* might still be scenarios in which you do so on a drunken whim. What gets rewarded in the toxin puzzle is not *trying to drink the toxin*. The relevant options here are not tryings, but intentions. When an agent decides ahead of the time to pursue some coarse of action, she adopts an intention to pursue that coarse. At least in part, this is a change in her beliefs and desires: she becomes more confident that she will act as she intends, and she will attach subjective cost to scenarios in which she acts otherwise. Perhaps other aspects of her mental state change as well. Perhaps her intentions even involve another basic propositional attitude – although it seems at least conceivable that a robot could plan ahead without having further basic attitudes.

So there are good reasons to think that in many decision problems, the relevant options are (propositions describing) intentions or decisions, rather than overt acts. Does this help with the certainty conditions?

Recall that according to (C1), a proposition A is an option only if the agent is certain that if she intended to make A true, then A would be true. Thus if one of your options is intending to go to Aleppo, (C1) would require you to be certain that if you intended to intend to go to Aleppo then you would succeed in intending to go to Aleppo. That doesn't sound plausible. Even if one can have intentions to have intentions – which is not obvious – we have no less infallible control over our intentions than over our actions.

Fortunately, (C1) was undermotivated anyway. (C2) looks a little more promising. Here the requirement is that merely by deciding to go to Aleppo you could rationally become certain that you intend to go to Aleppo. That doesn't sound too bad.

We noted that (C2) alone is not a useful recipe for an agent's decision mechanism to figure out the available options. Can we say more on which intention-describing propositions are available as options in a given context? Hedden [2012], drawing on Bratman [1987], suggests that a proposition of the type *S intends at t to* Φ is available as an option for S just in case the agent gives non-zero credence at t to the hypothesis that she can Φ. But should we really exclude intentions that are certain to fail? Couldn't the best way to achieve a goal be to aim for something impossible? One might also worry that there are hardly any values of Φ for which we are absolutely certain that if we intended to Φ then we wouldn't succeed. *Intending to win the lottery*, for example, will count as one of my present options.

There are other, more general problems about the present line of thought. First of all, it is not obvious that the result of a decision process is always a state of intention – especially if intentions are understood as special propositional attitudes. When you decide to go to Aleppo, you plausibly form an intention, but what about cases where you spontaneously raise a hand to greet a friend, or give way to other people on the street? These are choices. To be sure, they don't involve conscious deliberation, but they should at least in principle be captured by decision theory as a psychological model of

decision-making. Would an ideal robot need to form an intention whenever it moves around an obstacle?

Second, does forming an intention really make you rationally certain that you have formed the intention? If intentions are functionally individuated, this would mean that merely by forming an intention, an agent may become certain that she is in a state with a rather specific functional role. That is far from trivial. Again, must every decision-theoretic agent have the concept of an intention at all? (One may also worry about intentions with wide content, but I won't worry about that.)

Third, and most seriously, in many decision situations intentions seem too coarse-grained to do the work of options. Consider Jeffrey's marksman. The marksman presumably *intends to hit the target*. But that is way too unspecific to explain his movements. We might try to look for another intention that would be specific enough to qualify as the option he chooses. But this just leads us back to the main problem posed by the example: there seems to be no suitable more specific proposition the marksman intends to make true – especially if he is supposed to become certain that he is forming that intention.[9]

7 Options and outputs

The idea that options are intentions or decisions is on the right track. But I think we need to go a little further. To this end, I will first reconsider a decision-maker's options from an objective, external perspective. Here the decision-theoretic options present themselves as the possible immediate outcomes of a decision process. In the next section, I will consider how these outcomes should be represented in the agent's own doxastic system.

Imagine a simple decision-theoretic agent, with an internal representation of the actual state of the world, a representation of some goal state, and a motor system by which it can try to bring the actual state closer to the goal state – say, by moving around in its environment. If the agent's decision system determines that it should move to the left, a corresponding signal is sent to the motor system. We can imagine that a whole range of signals is available, controlling the angle and velocity of the movement. We might think of such a signal as representing that the agent moves to the left at angle x and velocity y, but that would not take into account that the resulting movement is not directly under the agent's control. For example, the velocity of the movement brought about by a given motor signal might depend on the inclination of the surface on which

9 In the analysis of perception, the analogous move to postulating intentions as options is to postulate *seemings* as perceptual inputs: what we learn through perception, on this account, is not that the external world is a certain way, but only that it seems to be a certain way. Here, too, the main problem is that propositions about seemings are far too unspecific to explain the change in our beliefs prompted by a perceptual experience.

the agent is standing, which may not be precisely known to the agent. Hence the choice of a given motor signal does not fix the resulting movements. If it is 50 percent likely that there is a wall immediately to the left of the agent, then a given signal to move to the left may be 50 percent likely to result in a movement to the left and 50 percent likely to result in no movement at all.

When the agent's decision system selects a motor signal, it therefore doesn't directly select a particular movement. What it needs to evaluate are not the possible effects of some overt movement, but the possible effects of sending the relevant signal. Given the agent's information about the world, each available motor signal determines a probability measure over possible movements, capturing the probability that the signal will make the agent move in a certain way. These probabilities are not fixed, but sensitive to the agent's information. By learning that there is a wall to its left, the agent can learn that signal α is unlikely to result in a movement in that direction.

The agent's options, then, are the possible motor signals it can produce. These are the direct outputs of its decision system, and the system should be designed so as to produce the optimal outputs: the outputs that maximize expected utility with respect to the agent's beliefs and goals.

If the agent is capable of strategic planning, there will be other outputs besides motor signals, adjusting the agent's beliefs, desires and commitments. It can also be useful to put further aspects of the agent's internal state under control of its decision system, such as her state of attention. (In principle, one could also adjust the agent's beliefs on decision-theoretic principles, so that she would always have beliefs that maximize expected utility.)

For moderately sophisticated agents, it might be useful to have a hierarchy of decision systems or decision steps, leading from the selection of fairly unspecific goals to detailed choices of how to achieve these goals. Return to Jeffrey's marksman. At some point, we may assume, he decided to shoot at the target. The result of this decision is a state of intention. In order to pursue the set goal, further decisions must be made. The marksman's cognitive system must integrate the available information concerning the wind, the gun, the terrain, the distance to the target etc. to determine a suitable stance – in effect, an adjustment of torques at the marksman's joints – that maximizes the probability of hitting the target. The output of that decision is still not a pattern of motor signals, for a given adjustment of torques can be brought about by countless muscle movements. A further, very low-level decision process is therefore needed to select muscle movements that efficiently bring about the desired torques. Ideally, any information acquired during this process should of course feed back to the higher-level decision states. If the most promising stance still comes with a very low probability of hitting the target, the marksman may decide not to shoot after all.[10]

10 The lower levels of the hierarchy just outlined form the topic of control theory, see e.g. [Todorov 2004].

On each level of the hierarchy, what is strictly speaking chosen is not the actual goal – say, a certain torque configuration – but an internal state of adopting that goal. The marksman's decision system does not have perfect control over the relevant torque positions, and thus must take into account various configurations that might result from a given decision. In addition, it is often easier to aim for an unspecific goal (a significant range of torque configurations) than to aim for more specific ones. The outputs of the decisions throughout the hierarchy will therefore not make any precise action, torque configuration or muscle movement certain, but only determine probabilities over these results.

Every decision process has a range of possible outputs. An optimal decision process should select the best output. From this perspective, there is nothing mysterious or difficult about the range of available options. The options are the possible outputs of the decision process. They are given as part of the design specification of the relevant process.

There is one major problem with this picture. If expected utilities are defined in something like the ways defended in [Jeffrey 1965], [Gibbard and Harper 1978], [Lewis 1981], [Skyrms 1984], [Sobel 1986] or [Joyce 1999], then the options must be elements of the algebra over which probabilities are defined. This is also needed for the agent to systematically update her beliefs about the likely effects of a given option. By learning that there is a wall to the left, the agent should learn that a certain signal is unlikely to produce a movement in that direction. If the signal is represented by some proposition in the agent's doxastic space, this episode can be modelled as a straightforward process of Bayesian learning. If the signal does not correspond to any proposition in the agent's doxastic space, it is quite unclear how the update is supposed to go.

So the outputs of a decision process should correspond to elements in the agent's doxastic system. Her options are not signals, but signal-describing propositions. But that seems odd. The relevant outputs of a decision process are complex physiological states. Must a decision-theoretic agent have sophisticated opinions about its own physiology? Worse, if we accept the certainty condition (C2), it would seem to follow that merely by making a choice, we should become certain of subtle facts about our own physiology. This is certainly false as a descriptive hypothesis, and it looks doubtful as a normative constraint.

Fortunately, there is a simple solution to these problems.

8 Inventing options

It is natural to think that if the possible outputs of a decision process are motor signals (say), then the corresponding propositions would have to be detailed descriptions of

[Todorov 2009] emphasizes the computational parallels to perception.

motor signals. But this isn't really necessary. The "signal-describing propositions" don't actually need to describe a signal. The marksman doesn't need to know – either consciously or subsconsciously – that his choice of a stance is a choice of such-and-such torque configurations. All we need are propositions that externally correspond to decision outputs in the sense that if the agent decides to make true an option proposition then the corresponding output is produced.

Imagine, for the sake of concreteness, that degrees of belief are defined over sentences in some language of thought. Let's say the language has sentences describing ordinary acts, states of the environment, and so on. There may also be sentences describing torque positions, beliefs, desires, intentions, etc. None of these are useful choices to serve as option propositions. Instead, when we design an ideal decision-maker, we should extend the agent's language of thought by 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. Each sentence is paired with a decision output – say, a motor signal – in such a way that when the agent makes a decision and "chooses" one of the sentences, then the corresponding output is produced. Due to this causal pairing, the new sentences might be regarded as "expressing" or "denoting" corresponding outputs. But from the internal perspective of the cognitive system, they do not carry information about physiology. If an agent gives positive credence to the hypothesis that there is no physical world, then making a decision does not require her to reduce that probability to zero, even though she will rationally become certain of some option proposition.

On the other hand, some aspects of the pairing must be reflected in the decision-makers beliefs. If the agent has no idea about 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. How could the agent acquire the "information" that X is likely to result in a certain movement, if 'X' doesn't express any genuine proposition about the world? Well, suppose 'X' is paired with a certain signal for moving to the left. When this output is produced, then the agent usually ends up moving to the left. From the agent's perspective, the truth of 'X' can be seen to correspond to movements in that direction. The agent can also learn how the resulting movement depends on the inclination of the ground, the presence of walls etc. So it is no mystery how the agent may come to assign a certain probability to the hypothesis that choosing 'X' will make her move to the left.

On a somewhat coarse-grained level, the agent might conceptualize her options in terms of their normal effects. If the agent knows that choosing 'X' normally leads to a movement to the left, she might conceptualize the relevant proposition as somehow related to such movements – as a particular "intention" or "decision to move to the left", perhaps. But this is not essential to the basic architecture. Crucially, what matters in decision making is the expected utility of the option proposition itself, not the expected utility of any

goal state that may be embedded in some conceptualization of that proposition.

The difference is especially salient under unusual circumstances where the agent knows that an option does not have its normal effects. In a well-known experiment on visual learning, people are given distorting goggles that shift everything they see to the left. When they are then asked to throw a basket into a ball, they (at least initially) have to consciously aim to the right of the basket. In a sense, they have to *intend to throw the ball to the right of the basket* in order to achieve the goal of getting the ball into the basket. (After several trials, their cognitive system adapts and the output previously conceptualized as aiming to the right of the basket gets re-conceptualized as aiming at the target.)

A complete theory of human decision-making should have more to say on these conceptualizations. The basic picture I have outlined does not require them. Even less does it require conceptualizations that are specific enough to distinguish between all available options. Option propositions don't need to be expressible in the agent's language. The agent doesn't need to have a language at all.

The primary objects of subjective probability are not sentences, but possible states of the world. To encode a probability measure over such states one might employ linguistic vehicles which in turn represent the possible states of the world, but there is no reason to postulate that the encoding must always take such a roundabout form. My proposal is therefore not really a proposal about a decision-maker's language of thought. The proposal is that the algebra of propositions over which subjective probabilities are defined should be extended by new elements corresponding to possible decision outputs.

If we focus on ordinary propositions about the decision-maker and her environment, or on quasi-linguistic, "conceptually structured propositions", we will not see the agent's options. Rational decision-making cannot be fully captured on that level. All we will see is something like Jeffrey's picture. When the marksman chooses a stance, he might become certain of various sentences or real-world propositions, but none of them will be specific enough to qualify as the option he chooses. His choice cannot be modelled as the selection of some proposition on this level. Instead, it corresponds to a redistribution of probabilities over a whole range of propositions. In one respect, these probabilistic consequences are all we need to know in order to evaluate the agent's choice, for she will hardly assign intrinsic value or disvalue to her option propositions. However, what is missing in that picture is a systematic account of where the available "lotteries" come from and how they are affected by the agent's evidence.

Remember Jeffrey's insight that our choices are part of the natural world and thus should be represented as ordinary propositions. The present account suggests that this is not quite right. From the decision-maker's perspective, her choices select propositions that stand outside the realm of ordinary physical propositions and that are only contingently and probabilistically related to physical propositions about the world. This might be

taken to explain or even vindicate the intuition that deliberate choices are "interventions", that they are not governed by the ordinary physical laws, and that they aren't suitable objects of real-world degrees of belief. I will not pursue these ideas further.

One other possible ramifications that may be worth pursuing concerns the treatment of "unstable" decision problems, such as the *Death in Damascus* scenario discussed in [Gibbard and Harper 1978]. Some, including myself, have argued that in these problems, decision theory cannot tell the agent what to do: it can only recommend a certain *state of indecision*, in which the agent assigns certain probabilities to the available acts (see [Arntzenius 2008], [Schwarz Unpublished]). In the present framework, we could still say that the agent should be undecided between the corresponding option propositions. But we could also model the recommended state of indecision as itself an option. After all, if options correspond to possible outputs of a decision process, and indecision is a possible output, then indecision should be an option.

9 Options and Actions

I have argued that from the perspective of decision theory (as a psychological model), an agent's options should be construed as primitive propositions corresponding to the possible outputs of her decision mechanism. Choosing an option goes hand in hand with become certain of the relevant proposition. Option propositions therefore satisfy condition (C2) from section 4.

If we tried to look for suitable propositions among sentences of English, or among ordinary ways things could be, we wouldn't find what we need. But when we design an agent's cognitive system we also get to design the algebra of propositions over which her probabilities are defined. And so we can simply add new propositions that do what we want.

Do we need to restrict the range of options from decision situation to decision situation, depending on the agent's information about her situation? Arguably not. Suppose you are absolutely certain that your arms are tied down, so that it would at best seem pointless to decide to raise your arm. Including the corresponding option proposition still doesn't seem to distort your decision problem. After all, you know that the relevant option will not succeed in bringing up your arm. So it probably won't come out as maximizing expected utility – unless there are other positive side-effects, as when you might get a reward for deciding to raise your arm, in which case we would go wrong by not including the option.

Of course, if we don't restrict an agent's option space, it will always be enormous – at least for agents remotely like us. There are indefinitely many intentions I could form at this moment, sentences I could think or utter, limb movements I could make. Any realistic cognitive architecture will have to use techniques for cutting down the options to

consider. But in this paper, my focus has been the specification of decision-theoretically ideal agents, without concern for computational realism or efficiency.

On the account I have defended, an agent's options are nothing like the acts we normally think of as her options. Only if the agent is absolutely certain that a given option will make true some act proposition can we identify her option with the corresponding act proposition. But usually the agent will not be certain about what act will result from a given option. There will be a whole range of possible results, associated with different probabilities.

Most decisions are ultimately about acts – about turning left or turning right, about going to Aleppo or staying in Damascus. On the picture I've painted, decision theory still makes predictions about an agent's behaviour. Recall that every option corresponds to a mental event with characteristic effects. Decision theory therefore predicts that an agent will behave in whatever way that mental event makes her behave in her given decision situation. For example, if the agent's arms are tied down but she doesn't know it, and it would further the agent's goals to raise her arms, then we predict that the agent will choose an option that would normally lead to a raising of her arm; the actual consequence of choosing that option will be a failed attempt at raising her arm. If the agent's arm isn't tied down, then circumstances are normal and we predict that she will indeed raise her arm.

Finally, we can return to the agent's *objective* options: to the things she can do. In section 2, I suggested that these could be analysed in terms of intentions: a possible act A is an option for an agent on a given occasion just in case the agent would perform A if she intended to perform A; alternatively: just in case some variation of the agent's intentional state would make her perform A. I said that I prefer the second, existential formulation. In the objective sense, it seems to me that the marksman *can* shoot at a given angle x and height y, even if it is not true that he would shoot in this manner if he intended to do so. Similarly, the marksman can intend to hit the target, even if he can't intend to intend to hit the target, so that nothing sensible can be said about what would happen if he intended to intend to hit. The existential analysis gets these cases right. Moreover, it is easily recast to connect with the proposed subjective conception of options: an act is an option for a agent in a given situation if some possible output of the agent's decision system would make her perform the act.

References

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

Frank Arntzenius, Adam Elga and John Hawthorne [2004]: "Bayesianism, infinite decisions, and binding". *Mind*, 113(450): 251–283

Jonathan Bennett [1988]: *Events and Their Names*. Oxford: Clarendon Press

— [1995]: *The Act Itself*. Oxford: Clarendon Press

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

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

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

Holly S. Goldman [1978]: "Doing the Best One Can". In A. Goldman and J. Kim (Eds.) *Values and Morals,* Dordrecht: Reidel, 185–214

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

Frank Jackson and Robert Pargetter [1986]: "Oughts, Options, and Actualism". *The Philosophical Review*, 95: 233–255

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

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

David Lewis [1969]: *Convention: A Philosophical Study*. Cambridge (Mass.): Harvard University Press

— [1979]: "Counterfactual Dependence and Time's Arrow". *Noûs*, 13: 455–476

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

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

Wolfgang Schwarz [Unpublished]: "Lost memories and useless coins: revisiting the absendminded driver". Under Review

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]

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

Emanuel Todorov [2004]: "Optimality principles in sensorimotor control". *Nature neuroscience*, 7(9): 907–915

— [2009]: "Parallels between sensory and motor information processing". *The Cognitive Neurosciences*

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