Options and Actions

Wolfgang Schwarz
Draft, 28 July 2014

1 Introduction

Decision theory says that a rational agent chooses acts that maximize expected utility. But even if we know an agent's beliefs and desires so that we can compute her expected utilities, this does not yield any predictions about what the agent will do, unless we 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, how humans make decisions. Or it could be a normative model, spelling out how one ideally ought to make decisions. Or it could be a constitutive model, implicitly defining what it is to be an agent with beliefs and desires. On each interpretation, we can't assume that the available acts are simply given as part of the specification of decision problems. 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, given that her decision mechanism isn't aware of the fact 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.

Reflection on this and other problems suggests 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 ordinary acts, although they are probabilistically related to acts. In the agent's cognitive system, they are best represented as primitive propositions.

2 What we can do

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, there is no point in specifying 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 try to figure out her 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 of 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 rough first stab, we might say that A is an option for an agent in a given situation iff some variation of the agent's psychological state would make her realize 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 gets things roughly right. On this account, A is an option for an agent just in case the agent *would* realize A if she *intended* to realize A. As a slightly weaker alternative, we might retain the existential form of our first stab and say that A is an option if some variation of the agent's intentional state would bring about A.

Neither of these will survive the discussion that follows, but it will be useful to have abbreviations for the two conditions. I will say that an agent in a given situation can^+ A just in case she would A if she intended to A; the agent can^- A just in case she would A under some variation of her intentional state.

One downside of these proposals is that they refer to a psychological state – intention – that is not part of the standard decision-theoretic inventory. Fortunately, we will see that there are other reasons to postulate intention-like states. These reasons will also

make a little clearer what exactly we need. For now, it is best to treat *intention* as a semi-technical term, with as yet not fully spelled out meaning.

The counterfactual analysis is usually discussed as an analysis of *ability* or the modal auxiliary *can*. As such, it is well-known to be a failure. We already saw that it doesn't capture the sense in which I *can't* eat the mince pie. It also doesn't 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 only I intended to play the piano, I would play). However, our topic is not the analysis of any pre-theoretic notion of abilities or options. For our topic these verdicts are exactly right. If there is no piano around, then playing the piano is not one of my decision-theoretic options.

What about another stock example: 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 not be in a coma)? Do we want to say that getting up is one of her options? Won't decision theory then falsely predict that she will get up – assuming she still has beliefs and desires? Again, that doesn't look like a serious problem. 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, our decision theoretic model describes the workings of an idealized, rational cognitive system. It doesn't apply to rocks or corpses or people in a coma. Decision theory is not meant to give correct predictions about the behaviour of these systems, so we can set aside the question how to delineate their options.

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 – say, from a utilitarian perspective –, then the relevant alternatives are precisely her decision-theoretic options. But I don't think that's generally true.

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. We often 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 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 options that figure in everyday normative evaluations can be more narrow than the options of decision theory. In other respects, they can be wider. In particular, decision-theoretic options (on our present understanding of decision theory)

do not include sequences of actions, unless the agent has some way of "binding" herself to the sequence. 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 options by our counterfactual analysis. From our decision-theoretic perspective, this is the correct verdict.[1] 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. No doubt there is 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, it might be useful to get clear on what counts as an option in decision theory, without presupposing a particular connection to our normative practice.

3 From acts to propositions

On the simple proposal from the previous section, an option for an agent in a given decision problem, or at a given time t, is a an act the agent can$^+$ (or can$^-$) perform – that is, an act A such that if the agent intended at t to perform A, then she would perform A. This puts a lot of weight on the concept of an act – more weight than the concept can sustain.

A first, and minor, complication is that agents generally realize only one of their options. The other options therefore can't be actual, token events. At best, they would have to be merely possible events. But if the decision-maker faces an actual choice between several options, it would seem to follow that these merely possible events, despite being merely possible, actually exist.

More importantly, decision theory requires us to trace the same options across a wide

1 Doesn't that make the professor irrational, so that he falls outside the purview of decision theory? I don't think so. The professor might be "irrational" in several respects, for example, insofar as his choices in life are not the choices he wishes he would make, but he is not necessarily irrational in the very restricted sense of decision theory. His actions are rationalized by a perfectly coherent preference order.

range of hypothetical circumstances. It is not at all clear that this is even possible with ordinary token acts, and whether it yields sensible results. Suppose while preparing an omelette you 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. Could you have performed *this very act* if the egg had been good? One might intuit that the answer is no. However, to compute the expected utility of an act, we have to consider its outcome in all possible states of the world. In particular, we have to consider the outcome of performing α in situations in which the egg is not rotten. So do we have to consider impossible situations in which an act of adding a rotten egg is an act of adding a good egg?

These sorts of difficulties go away if we identify options not with acts, but with act-describing propositions (or properties, or act types). The act of adding an egg may be identical to the act of adding a rotten egg, but the propositions *that you add an egg* and *that you add a rotten egg* are uncontroversially different. Only the former provides a reasonable characterisation of the option you chose.[2]

In decision theory, the idea that options should be construed as propositions was prominently defended by Richard Jeffrey in his *Logic of Decision* [Jeffrey 1965]. Earlier models, such as that of [Savage 1954], had drawn a sharp distinction between acts and *states of nature* which serve as objects of ignorance and belief. On Jeffrey's account, this distinction is dropped: "the human agent is taken to be part of nature and his acts are thus ingredients in states of nature" [Jeffrey 1992: 226].

Jeffrey's proposal opened up the possibility of appealing to probabilities of options, probabilities conditional on options, or probabilities of complex propositions involving options. In Jeffrey's own "evidential" decision theory, expected utilities are computed by conditionalizing the agent's probability function on the relevant option proposition. An analogous form of "causal" decision theory is defended in [Joyce 1999]. Other formulations of causal decision theory appeal to conditionals with options in the antecedent ([Gibbard and Harper 1978], [Lewis 1981]). Probabilities over options are also put to use, for example, in Skyrms's [1990] model of rational deliberation.[3]

Another advantage of identifying options with propositions is that we can then allow for options that don't correspond to intuitive act types. In Kavka's [1983] toxin puzzle, we can say that you don't just face a decision between *drinking the toxin* and *not drinking*

2 Of course, we can still go on and call act-describing propositions 'acts' – as is common in decision theory. The point is that decision-theoretic options are not acts in the ordinary sense of an act. More reasons for individuating acts as propositions can be found in [Bennett 1988] and [Bennett 1995].

3 Not everyone is convinced by these applications. Some hold that deliberation is incompatible with having credences about one's choices. 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.

the toxin, but also between *intending to drink the toxin* and *not intending to drink it*. I will return to this point in section 6, where we will consider the view that options are always intention-describing propositions rather than act-describing propositions.

If we model options as propositions, we run into the verbal complication 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.

We can easily restate the preliminary definitions of the previous section in terms of propositions as options. An option for an agent at time t is a proposition the agent can$^+$ (or can$^-$) make true at t.

The main defect of this proposal is that it doesn't take into account the agent's epistemic perspective: what she knows or fails to know about what she can do. Before we turn to that, I will briefly address a smaller problem concerning the specificity of options. On the present account, every proposition entailed by an option is itself an option: if you can make it true that you raise your arm, then you can also make it true that you move a limb. However, there are reasons for excluding these unspecific propositions and focus on the most specific propositions an agent can make true.

For example, consider a situation in which you are punished for intentionally raising your arm but rewarded for unintentionally raising it. Suppose you have no way of deliberately 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. In fact, your only relevant option is the more specific *intentionally raising your arm*, which has low expected utility.[4]

So let's add a specificity clause. An option for an agent at a given time is a proposition the agent can$^+$ (or can$^-$) make true at the time that is not entailed by a stronger proposition the agent can$^+$ (or can$^-$) make true.

4 The decision theoretic literature on this point has mostly focused 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 the (exclusive) 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. However, [Sobel 1983] 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 specifically in causal decision theory is that the usual rules for evaluating causal expected utility become problematic if the relevant options are very unspecific. Subjunctive conditionals, imaging functions, and conditional chances are hard to evaluate if the "antecedent" proposition is unspecific.

4 Epistemic access

The definition we've arrived at can be seen as a fleshed-out version of Lewis's definition in "Causal Decision Theory", there presented without argument:

> 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]

But that can't be right. 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 both can$^+$ and can$^-$ make true that proposition.) 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. But given your state of uncertainty about the switch, it would be wrong to count the latter as one of your options. Suppose you give some probability to the hypothesis that pushing the button would call the 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.

The problem is that so far our definitions are not sensitive to the decision-maker's epistemic state. While it is true that you can$^+$ turn on the light by pushing the button, this is not something you know. Intuitively, if we're interested in decision theory as a psychological model, then what counts as an option should not be sensitive to external facts of which the agent isn't aware. The agent's decision module has to make do with whatever information it has.

A popular move in response to these worries is to define options not in terms of what the agent can *in fact* make true, but in terms of what we she *believes* she can make true – where 'belief' means credence 1. Following that line of thought, we might say that a proposition is an option for an agent at a time iff it is a maximally specific proposition of which the agent is certain that she can$^+$ make it true (that is, a proposition of which she is certain that she can$^+$ make it true that is not entailed by a stronger proposition of which she is certain that she can$^+$ make it true). Let's call this *proposal 1*.

(We have to use *can*$^+$ here, not *can*$^-$. For suppose you see two buttons, and know that one of them operates the lights while the other calls room service, but you don't know which does which. Then you know that you can$^-$ make true the proposition that you press one of the buttons and turn on the lights. However, in an adequate representation of your decision problem, none of the options should entail that proposition.)

One problem with proposal 1 is that it goes beyond what its motivation requires. 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 certainty. Compare the agent's degrees of belief and desire. These are presumably accessible, in the relevant sense, to her decision-making mechanism – even if the agent is not absolutely certain about her beliefs and desires.[5]

Proposal 1 also seems to run into the issue I mentioned in the introduction: it allows 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 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.

The most serious problem with proposal 1, however, is that if we look at ordinary decision situations, it is hard to find any interesting proposition of which the agent is absolutely certain that she can$^+$ make them true. For example, I am not absolutely certain that my muscles will keep functioning during the next few seconds; so I am not absolutely certain that I can$^+$ raise my hand. Is there anything 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.

Proposal 1 is not the only way of taking into account an agent's epistemic perspective. Jeffrey [1968] discusses the following condition: a proposition qualifies as an option only if it would be rational for the agent to become certain that the proposition is true by making a choice. In the hotel room example, this excludes *turning on the lights by pushing the button*, since it would not be rational for you to become certain that the lights will go on merely by deciding to push the button.

The assumption that motivates Jeffrey's condition is that decision-making provides "knowledge without observation", as Anscombe said. Having decided to order waffles, 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. This intuition is supported by the models of deliberation outlined in [Skyrms 1990], on which rational deliberation goes along with becoming certain that the

5 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.

eventually chosen proposition is true.

There are different ways of incorporating Jeffrey's condition in a definition of options. Most simply, we might say that an option for an agent is a most specific proposition of which she might rationally become certain by making a choice. Let's call this *proposal 2*.

While it is a little better motivated than proposal 1, proposal 2 also has some obvious problems. For one thing, as it stands it offers little guidance for a mechanism aiming to find a decision-maker's options. How is the agent's cognitive system supposed to determine what it would be rational for the agent to believe through making a choice?

We also still face the problem that decision situations may not come with sufficiently specific propositions that satisfy the definition. 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 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 proposal 2, that it would be rational of her to become *reasonably confident* that the proposition is true by making a choice?

Well, return to the hotel room example. 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.

To be sure, 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 isn't just hard to find suitable propositions of which the agent is absolutely certain, it is even hard to find propositions of which she is reasonable confident. Jeffrey considers the case of a marksman aiming at a distant target.

> [The] marksman may have a fairly definite idea of his chances of hitting [the target], e.g., he may have degree of belief 0.3 in the proposition H that he

> will hit it. The basis for his belief may be his impressions of wind conditions, quality of the rifle, etc.; but there need be no reason to suppose that the marksman can express the relevant data. [Jeffrey 1968: 37]

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 the marksman can$^+$ make true, nor something of which he beliefs that he can$^+$ make it true, nor something of which he ought to become certain merely through making a choice.

Jeffrey's solution, presented in [Jeffrey 1965: sec. 11.9] and [Jeffrey 1968], 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 options are given by the partition $\{H, \neg H\}$ with associated probabilities 0.3 and 0.7. Instead of making him certain of anything, 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 not hitting*, *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. Formally, the options look a lot like the "lotteries" or "prospects" or "mixed strategies" in earlier models of decision, except that the "pure options" over which the mixtures are defined are partitions of ordinary propositions.

All this is very elegant, but remember our topic. We are looking for a recipe that tells us how to catalogue an agent's options in a concrete decision situation. On Jeffrey's account, this means that we need to determine which partitions of propositions with which probabilities are available. Why does the marksman have options on the partition $\{H, \neg H\}$, as opposed to infinitely many other possible partitions? Why does one of his options correspond to the probabilities 0.3, 0.7, and none to $0.1, 0.9$? Clearly the reason has something to do with the marksman's evidence. As Jeffrey notes, the numbers reflect the marksman's information about the wind, the quality of his rifle, etc. So how does this work? How does the information available to an agent determine a set of Jeffrey-style options?

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. Loosely speaking, decision theory asks you to pay special attention to the (epistemically or subjunctively) most probable worlds at which a given option is realized. Thus, if *going to Aleppo in the evening* is one of your options, we should ask what is likely to happen if you go to Aleppo. Now 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 she is still about to receive information potentially relevant to her decision. 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.[6]

The possibility of substantive gaps between the time of decision at the time of acting on the decision also strengthens our earlier arguments against identifying options with the relevant acts. Your decision to go to Aleppo in the evening by no means makes it certain that you will go. You might find out that the road is blocked, or that horses are too expensive. These possibilities should be taken into account – as they would not if you simply considered the expected utility of going to Aleppo.

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

Now let's return to the above idea than an agent's options are generally intentions. Following our discussion in section 3, we better say that they are intention-describing propositions. Do these propositions satisfy our certainty conditions?

Remember the two proposals from section 4. Proposal 1 required that the agent must be certain that she can$^+$ make true her options. Accordingly, if *intending to go to Aleppo*

6 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; it is useful in deterrence cases and the decision puzzles of [Arntzenius et al. 2004]; it might also enable useful social institutions and conventions.

is one of your options, you would have to be certain that you can$^+$ intend to go to Aleppo. That is, you have to be certain that *if you intended to intend to go to Aleppo then you would succeed in intending to go to Aleppo.* This 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.

In this context, conditions in terms of *canx* look more plausible than conditions in terms of *can*. Recall that an agent can$^-$ make true a proposition A at a time t iff there is some variation of her intentional state that would make A true. If A itself describes a variation of the agent's intentional state, this condition becomes trivial.

What about proposal 2, on which options are the most specific propositions of which it would be rational for the agent to become certain merely by making a choice. In the Damascus example, this requires that at the end of your deliberation you could become rationally certain that you intend to go to Aleppo. That doesn't sound too bad.

We noted that proposal 2 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 *S intends at t to* Φ is available as an option for S just in case the agent gives non-zero credence to the hypothesis that she can$^+$ Φ at t.

I am not convinced that we should 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. Also, what about values of Φ for which the agent can't even intend to Φ? But let's set these worries aside.

There are others. For one thing, 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. (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.[7]

7 Invention options

I think the idea that options are intentions or decisions is on the right track. But I want to flesh it out a little differently. Doing so will take two steps.

The first is to think of options as the immediate (possible) outcomes of a decision process. Here we look at the agent from the perspective of an engineer or cognitive scientist, inspecting the internals of the agent's decision mechanism. Taking into account the agent's own perspective on her options will be the second step.

So imagine we look at the functional specification of an agent's cognitive architecture. If the architecture follows the principles of decision theory, what will we find? Put differently, how would we go about designing such a specification?

A very simple architecture might apply principles of decision theory to directly alter the state of the agent's motor system – her muscles, say – based on her beliefs and desires. The direct output of the decision process would here be motor signals. If the agent is capable of strategic planning, there will be other outputs, adjusting the agent's beliefs, desires and commitments. It can also be useful to adjust further features of the agent's internal state on the basis of expected utility calculations, e.g. her state of attention. We could, if we wanted, even implement updates of the agent's beliefs on decision-theoretic principles, so that she'd always have beliefs that maximize expected utility – although this would probably be a bad idea if we want the agent to successfully interact with her environment.

For moderately sophisticated agents, it might be useful to implement a hierarchy of decision processes, leading from the selection of fairly unspecific goals to detailed choices of how to achieve the goals. Consider the marksman's cognitive system. At some point, we may assume, the marksman decided to shoot at the target. The result of this decision is a state of intention. In order to pursue this goal, further decisions must be made.

7 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 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 needed to select muscle movements that efficiently bring about the desired torques.[8] Ideally, any information acquired during this process should of course feed back to the higher-level decision states. If even the most promising stance comes with a low probability of hitting the target, the marksman may decide not to shoot after all.

The situation is further complicated by the fact that the marksman's decision mechanisms have imperfect control not only over whether he will eventually hit the target, but even over his muscle movements and torque configurations. There is always noise in biological systems. Moreover, it is often easier to aim for an unspecific goal (say, a significant range of torque configurations) than to aim for more specific ones. As a consequence, the outputs of a decision process throughout the hierarchy will not make any precise action, torque configuration or muscle movement certain, but only determine certain probabilities over these goals.

Every decision process has a range of possible outputs. An optimal decision process should select the best output – that is, the output that maximizes expected utility with respect to the system's information and goals. From this perspective, the options whose expected utility should be considered are therefore simply the possible outputs of the relevant decision process.

There is nothing very mysterious or difficult about the set of possible outputs of a decision process. It is part of the design specification of the process. Consider a low-level process that chooses muscle movements. Here the possible outputs are motor signals. Which signals are available is a matter of how the system is built.

But things are not quite that simple. If a decision process is to compute the expected utility of an option, it needs to somehow represent the option, and it needs to represent it in such a way that it can sensibly evaluate the utility of the option under various hypotheses about the world. In other words, we need to translate the possible outputs of a decision into output-describing propositions in the agent's doxastic space. The agent can then consider, for example, the probability of various consequences conditional on one of these propositions.

It is natural to think that if the available outputs are (say) motor signals, then the corresponding propositions would have to be detailed descriptions of motor signals, in physical or functional terms. But this isn't really necessary. The "output-describing

8 The lower levels of this hierarchy form the topic of control theory, see e.g. [Todorov 2004]. [Todorov 2009] emphasizes the computational parallels to perception.

propositions" don't actually need to describe an output. 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 that 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, I suggest, 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 any information about the relevant physical or functional facts. 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 decision-maker 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. Fortunately, the needed information is not hard to come by. Suppose 'X' is paired with the output of what is intuitively a decision to raise the agent's left hand. When this output is produced, then almost always the agent's left hand is about to go up – except in unusual situations in which her arm is tied down etc. Having witnessed this correlation, the agent can reasonably infer that choosing 'X' will make it likely that her arm will go up (unless it is tied down etc.).

In cases like this, the agent might conceptualize her option as a decision or intention to raise her arm. In general, if the agent knows that choosing an option O normally gives rise to the truth of a proposition A, then it makes sense for her to conceptualize O not as a bare tag, but as somehow related to or embedding A. But this is not essential to the basic architecture. What's important for decision making is not the expected utility of the embedded proposition A, but the expected utility of the option O itself.

The difference is especially salient under unusual circumstances where the agent knows that O does not go along with A. When people are given distorting goggles that shift everything they see to the left, and are asked to throw a basket into a ball, they (at least

initially) have to aim towards 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. It doesn't require that option propositions are expressible in the agent's language. Decision-theoretic processes can be implemented in agents that don't have a language at all. This is why it is better to define subjective probabilities (and utilities) not over sentences in an agent's language, but over an abstract algebra of possibilities. My proposal is that this algebra 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. What we will see is something like the picture drawn by Jeffrey. When the marksman chooses a stance, he might become certain of various things, 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 real-world or linguistic proposition. 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 the agent 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. If you come to believe that your arm might be tied down, you suddenly no longer have an option that confers 0.999 degree of belief on the proposition that your arm will go up. If the marksman observes that the wind has become stronger, her best option may only be associated with a probability of 0.2 of hitting the target. In the present model, what's going on here is that the agent's beliefs about the likely consequence of the very same options change. All uncertainty is located in the decision-theoretic "states of nature". We don't need a separate story about the probabilities in an agent's options.

Remember Jeffrey's key insight, that our choices are part of the natural history of the world and thus should be represented by 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.

Another possible consequence that I will only mention concerns the treatment of unstable decision problems. The most famous example (from [Gibbard and Harper 1978]) concerns a choice between going to Aleppo and staying in Damascus. But the story is a little more complicated than the one we've discussed so far. The complication is that at one of the two places, Death awaits you, depending on where he foresaw that you will be. Death is a very good predictor, so if you go to Aleppo, it is likely that this is where he is waiting, in which case it would have been much better to stay in Damascus. Similarly, if you stay in Damascus, it is likely that Death is waiting for you there, in which case it would have been much better to go to Aleppo. Some, including myself, have argued that in cases like these, decision theory can't tell you what to do. It can only recommend a certain *state of indecision*. You should be undecided between staying in Damascus and going to Aleppo, with a certain probability assigned to either coarse of action. In the present framework, we could still say that you should be undecided between two options – intuitively, between intending to go to Aleppo and intending to go to Damascus. But there is another way of modelling the case: we could say your recommended state of indecision is itself an option. After all, options in general map onto possible outputs of decision processes associated with probabilities for ordinary propositions.

8 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. So these propositions satisfy proposal 2 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. However, including the corresponding option proposition 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.

References

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

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

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

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

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

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