Transformative experience and decision theory[*]

Richard Pettigrew

I have never eaten Vegemite—should I try it? I currently have no children—should I apply to adopt a child? In each case, one might imagine, whichever choice I make, I can make it rationally by appealing to the principles of decision theory. Not always, says L. A. Paul. In *Transformative Experience*, Paul issues two challenges to decision theory based upon examples such as these [Paul, 2014]. I will show how we might reformulate decision theory in the face of these challenges. Then I will consider the philosophical questions that remain after the challenges have been accommodated.

1 Deliberative and justificatory decision theory

The subject matter of decision theory is decision problems. We model a decision problem in which an agent must choose between a range of acts as follows:

(i) *Acts* $\mathcal{A}$ is a set of propositions each of which describes a different possible act that our agent might perform and states that she does in fact perform that act.

(ii) *Preferences* $\preceq$ is a preference ordering on the set of acts $\mathcal{A}$.

(iii) *States* $\mathcal{S}$ is a set of propositions each of which describes a different possible state of the world; they are mutually exclusive and exhaustive.

(iv) *Utilities* u is a function that takes a conjunction of the form $A \wedge S$, where A is in $\mathcal{A}$ and S is in $\mathcal{S}$, and returns the utility that the agent would obtain were that conjunction to hold: that is, the utility she would obtain if she were to perform that act in that state of the world. Conjunctions of the form $A \wedge S$ are called *outcomes relative to* $\mathcal{A}$ *and* $\mathcal{S}$.

(v) *Credences* p is a subjective probability function defined on subjunctive conditionals of the form $A > S$, where A is in $\mathcal{A}$ and S is in $\mathcal{S}$: this is the conditional *If I were to perform act A, then S would be the state of the world*. Conditionals of the form $A > S$ are called *dependency hypotheses relative to* $\mathcal{A}$ *and* $\mathcal{S}$.

(vi) *Expectations* We define the *expected utility of A relative to p and u* as follows: $E_p(u(A)) = \sum_{S \in \mathcal{S}} p(A > S)u(A \wedge S)$.

[*] I'm extremely grateful to Samir Okasha, Ben Levinstein, Jason Konek, Greg Wheeler, and Anya Farennikova for very helpful discussion of the proposals mooted in this paper; and, of course, to Laurie Paul for many wonderful, fascinating conversations. I was supported by an ERC Starting Researcher grant 'Epistemic Utility Theory: Foundations and Applications' during my work on this paper.

For many decision theorists, the only attitudes of the agent that have substantial psychological reality are their preferences. While these decision theorists accept that a rational agent has a utility function and a credence function, they contend that this says no more than that that those functions together *represent* the agent's preference ordering in the sense that the ordering of the acts by their expected utility relative to those functions matches the ordering of the acts given by the agent's preference ordering: that is, $E_p(u(A)) \leq E_p(u(B)) \Longleftrightarrow A \preceq B$. They maintain that rationality imposes conditions on the agent's preference ordering and they show that any ordering that satisfies those conditions is representable by a unique probabilistic credence function and a utility function that is unique up to affine transformation. Thus, for these decision theorists, decision theory is concerned only with the rationality requirements that govern preferences, and the ways in which we can represent agents who satisfy those requirements. It is not concerned with how we might deliberate about which preferences to have nor with how we might justify those preferences. On this view, I can justify choosing an act by noting that I prefer it to all others. But, if you ask me to go further and justify those preferences, there is nothing I can say. I cannot appeal to my credences and utilities. I have those credences and utilities in virtue of having the preferences I have; so they cannot justify those preferences. We might call this the *preference-first conception of decision theory*.

However, these justificatory and deliberative tasks are clearly important. When you ask me to justify my preferences, I do not fall silent in the way predicted by the preference-first conception. Instead, I appeal to my credences and my utilities and the expected utility values for acts relative to them. That is, I treat credences and utilities as psychologically real and capable of justifying the preferences I have. Thus, there is an alternative account of decision theory on which at least part of its job is to say how I can appeal to credences and utilities to justify my preferences or to set those preferences in the first place. This sort of decision theory can be used by an agent deliberating about a choice she must make; and it can be used by an agent after she has made her choice at the point at which she needs to justify it. When she is deliberating, she attends to her credences and her utilities, she uses them to calculate her expected utilities, and she chooses an act with maximal expected utility. When she is justifying her choice, she demonstrates that it maximises expected utility relative to her credences and utilities. We might call this the *deliberative conception of decision theory*. It is to this conception that L. A. Paul addresses her challenges. These challenges do not affect the preference-first conception: that is, they do not provide counterexamples to the conditions that rationality is taken to impose on an agent's preference ordering.

2 Epistemically transformative experience

The first challenge arises from the existence of epistemically transformative experiences (ETEs). Recall: ETEs are those that teach you something about the phenomenal character of an experience that can only be learned by having that experience.

Suppose I face a choice between a range of alternative acts; and suppose that one of those acts has a possible outcome that involves an ETE. For instance, suppose I have never tried Vegemite. And suppose I must choose whether or not to accept a bet on a coin flip that gains me a piece of toast spread with Vegemite if the coin lands heads, and loses me £1 if it lands tails. Now, if I am to use decision theory to help me deliberate about what to do, or to justify my decision once it is made, I must at least have access to my credences over the relevant dependency hypotheses and to the utilities I assign to the outcomes of the available acts. After all, I must use them to calculate the expected utilities of the available acts. The problem is that, on one plausible formulation of the decision problem I face, I do not have access to these utilities. On this formulation, $\mathcal{A} = \{\mathit{Bet}, \mathit{Don't\ bet}\}$ and $\mathcal{S} = \{\mathit{Heads}, \mathit{Tails}\}$. Now, while I know the utility I assign to the outcome *Don't bet* (it is just the status quo) and the outcome $\mathit{Bet} \wedge \mathit{Tails}$ (it is just the disutility of losing £1), I don't have access to the utility I assign to $\mathit{Bet} \wedge \mathit{Heads}$. On this latter outcome, I gain a piece of toast spread with Vegemite, so a large part of what will determine the utility I assign to it is the phenomenal character of the experience of eating Vegemite; and this is something to which I lack access at the time the decision must be made, since the experience is an ETE.[1]

This is Paul's first challenge to the deliberative conception of decision theory. A natural response is to reformulate the decision problem in question by reframing the uncertainty about the utility function as uncertainty about the world. I will call this the *redescription strategy*. Thus, instead of taking the set of possible states of the world to be $\mathcal{S}$, we instead take it to be $\mathcal{S}'$, a fine-graining of $\mathcal{S}$, where the fine-grained possible states of the world specify not only how the world is, but also what my utility function is over the outcomes relative to $\mathcal{A}$ and $\mathcal{S}$. Thus, let u_1, ..., u_n be the utility functions I might take have over the outcomes relative to $\mathcal{A}$ and $\mathcal{S}$. Then we reformulate the decision problem by taking the set of possible states of the world to be:

$$\mathcal{S}' = \{S \wedge \mathit{My\ utility\ function\ is\ } u_i : S \in \mathcal{S} \text{ and } i = 1, \ldots, n\}$$

And we define the utility of the outcomes relative to $\mathcal{A}$ and $\mathcal{S}'$ as follows: for each $i = 1, \ldots, n$,

$$u(A \wedge S \wedge \mathit{My\ utility\ function\ is\ } u_i) = u_i(A \wedge S)$$

Doing this solves the original problem of epistemically inaccessible utilities. But it requires me to have credences over a new set of dependency hypotheses, such as, $\mathit{Bet} > (\mathit{Tails} \wedge \mathit{My\ utility\ function\ is\ } u_i)$. Is this a problem? I think not. Firstly, the utility hypotheses simply specify numerically the utility the agent obtains at each outcome. So you don't need to know the possible phenomenal characters of the experiences that you will have at each outcome in order to know what the possible utility hypotheses are to include in $\mathcal{S}'$. You simply need to know the possible values of your utility function—and you do know that, since possible utility values are all real numbers. Secondly, once you've set out the various possible utility hypotheses, they are simply empirical hypotheses about which you can accumulate evidence in the usual way.

This, then, is the redescription strategy: faced with uncertainty about the utilities I assign to possible outcomes of available acts, I simply fine-grain the possible states of the world to include the various possible utility hypotheses about which I am uncertain; my utility function over the possible outcomes of the acts relative to these states of the world is then accessible to me and I can quantify my uncertainty about my utility function over the original outcomes using my credence function over the new dependency hypotheses. Does this answer Paul's challenge? Paul thinks not.

Paul's concern is that, when I am uncertain of my utility function, the redescription strategy exhorts me to make decisions based on my credences over hypotheses about my own

[1] In fact, Paul allows that an agent might make such a decision rationally if, for instance, she has a strong desire to have a new experience. I will be concerned with the cases in which there are not these other motivating desires.

utility function; but, when an ETE is involved, these credences must be based in turn on statistical evidence that summarises what *other people* report about *their* utility functions; by assumption, none of the evidence can be about *my* utility function. Why is this a problem? We are accustomed to using statistical evidence based on facts about other people in order to make decisions about ourselves: I choose to exercise because statistical evidence based on data from other people suggests that it will improve my health. Why is it different when the statistical evidence bears on the hypotheses about my own utility function? Paul's concern is that, by basing our credences over the utility hypotheses solely on data about others and not about me, I threaten the *authenticity* of my decision.

> [W]e also want to choose authentically, that is, we want to choose in a way that is true to ourselves, in a way that involves our self as a reflective, deliberating person, choosing after assessing our preferences from our first-personal point of view and then living with the results. [Paul, 2014, 128]

Thus, Paul's worry stems from existentialist concerns. Elsewhere, she compares making decisions based on statistical evidence that summarises the experiences of others to making decisions based on the dictates of a putative morality whose values you do not fully internalise as your own, or based on the dictates of a group of which you are a member but with whom you share few values. The latter cases are, of course, the existentialist's archetypes of inauthentic action. And the existentialist's concern is that they alienate you from your decisions.

I contend that these cases are not analogous. While it may be true that I will be alienated from my decision if I simply defer to the dictates of morality or my social group and studiously ignore my own utilities, it is not true in the cases of decisions made on the basis of credences that quantify uncertainty about my own utility function. The reason is that, although in the latter case my evidence does not include facts about my own utility function, it nonetheless provides evidence that supports credences in propositions concerning my own utility function. Indeed, my purpose in collecting statistical evidence based on the utilities of others is precisely to try to overcome the epistemic barrier to knowing my own utilities in the case of ETEs.

Suppose I read a survey that reports that, amongst a large randomly selected sample from a population of which I am a member, 70% of respondents assigned a high utility to being a parent, while 30% assigned a low utility. In the absence of other evidence, I might reasonably assign a credence of 70% that *I* will assign high utility to being a parent. But notice what would happen were I then to learn that all members of this sample population had drastically different utilities from mine in every other sphere. In this situation, I would reasonably abandon the credences I had assigned, because I would have learned that the utilities assigned by this population were not a good indication of my own. What this shows is that, when I use statistical evidence to set my credences about my own utilities and then use these credences to make a decision, I am not simply deferring to the majority opinion in a way that renders my decision inauthentic. Rather, I am using the opinions of others as evidence about my own utility function. Put another way: I attend to the opinions of others not because I wish to follow the majority decision, but rather because I want to find out about myself.

Thus, while facts about my utilities may not be contained in the evidence, my utilities are nonetheless a key ingredient in my deliberation in a way that is very different from the case in which I simply defer completely to morality or to the mores of my social group. There is a difference between being certain of my utilities and ignoring them in favour of acting in accordance with moral laws or group decisions, on the one hand, and being uncertain of my utilities and using all available evidence in order to predict them and then acting in accordance with my best possible predictions about them, on the other. One way to see the difference is to note that, when I make a decision based purely on the demands of morality, the utility function that I use in the expected utility calculation is not my own—it is rather an objective value function that encodes the demands of morality. Thus, when I choose, I do not choose the action that I expect to have highest utility by my lights; I choose the act that I expect to have highest utility by the lights of objective morality. That is alienating. The same is true if I use the utility function of my social group when I make my decision, rather than my own utility function. Again, that is alienating. On the other hand, when I choose between acts that might give rise to ETEs, I choose the act that I expect to have highest utility by *my* lights, even though I'm uncertain about what those lights are. In this case, it seems to me, the decision need not be alienating.

3 Personally transformative experience

Paul's second challenge arises from the existence of personally transformative experiences (PTEs). Recall: PTEs are those that lead you to change what you value and to what extent.

Suppose I face a choice between a range of acts; and suppose that one of those acts has a possible outcome that involves a PTE. For instance, suppose I must choose whether or not to become a parent for the first time. Thus, in this decision problem, $\mathcal{A} = \{Adopt, Don't\ adopt\}$. And let's assume that there is no uncertainty in the world—if I choose to adopt, I will adopt; if I choose not to, I won't. Thus, on a natural formulation of the decision problem, there is just one possible state of the world in $\mathcal{S}$. Thus, the outcomes relative to $\mathcal{A}$ and $\mathcal{S}$ are just *Adopt* and *Don't adopt*. In the former, I become a parent; in the latter, I don't. And let us say, very crudely and ignoring the fact that becoming a parent may be an ETE as well as a PTE, that I know that, if I become a parent, I will come to value time spent with friends less and time spent with family more. What's more, I know that the outcome *Adopt* involves a lot more time with family and a lot less with friends, while this is reversed in *Don't adopt*. Thus, my current utility for *Adopt* is lower than for *Don't adopt*; and this is true of my future utilities as well in the outcome in which I don't adopt. But, if I do adopt, then my future utilities will be the reverse of my current utilities: I will value the family time entailed by *Adopt* more than the time with friends entailed by *Don't adopt*. How, then, am I to make this choice? This is Paul's second challenge to decision theory: it is ambiguous in cases in which my utilities change over time.[2]

A natural response is this. First, introduce the notion of a *local utility function*: my local utility function at a given time is the function that measures how much I value outcomes at that time. Next, let us demand that, as in the previous section, the possible states of the world are fine-grained enough to include a specification of the relevant facts about my utilities: in this case, where my utilities may change over time, that will include a specification of my local utility function for each time during my life in that state of the world. Now, let us fix attention on a particular outcome that results from conjoining an act with a state of the world

[2] Again, Paul allows that such a decision may be made rationally if the agent has some other motivating desire, such as a desire to have an heir. Again, I will be restricting attention to the case in which there is no such other motivating desire.

that is fine-grained in that way: thus, this outcome will include a specification A of the act that is performed and a specification S of how the world is (call the conjunction $A \wedge S$ its *worldly component*); and it will include a specification of my local utility functions (call this its *utility component*). Let us simplify and assume that, in this outcome, there are just finitely many moments in my life, $t_1, \ldots, t_n$. And let $lu_{t_1}, \ldots, lu_{t_n}$ be my local utility functions at those moments. Thus, the outcome in question is:

$$A \wedge S \wedge \bigwedge_{i=1}^{n} \mathit{My\ local\ utility\ at\ } t_i \mathit{\ is\ } lu_{t_i}$$

Then we might say that the utility function to which I appeal when I make a decision at one of those moments (say t_i) is my *global utility function at t_i* (which we write u_{t_i}), where the utility that u_{t_i} assigns to the outcome just described in some way aggregates the local utilities assigned to the worldly component of that outcome ($A \wedge S$) by each of the local utility functions $lu_{t_1}, \ldots, lu_{t_n}$. The natural means of aggregation in this case is the weighted sum. Thus, my global utility at t_i for the outcome described above is determined by a series of weights $\beta_{t_1}, \ldots, \beta_{t_n} \geq 0$. These specify how much each of my local utilities for the worldly portion of that outcome contribute to the global utility for the whole outcome, which is the utility I use when I make a decision or justify it once made. So, for each $j = 1, \ldots, n$, the weight β_{t_j} specifies the extent to which I take into account the values I had or have or will have at time t_j. Thus:

$$u_{t_i}(A \wedge S \wedge \bigwedge_{i=1}^{n} \mathit{My\ local\ utility\ at\ } t_i \mathit{\ is\ } lu_{t_i}) = \sum_{i=1}^{n} \beta_{t_i} lu_{t_i}(A \wedge S)$$

This solves the formal problem. That is, it describes a formal framework that can be used to make decisions in cases in which (local) utilities might change. Yet in doing so, it does little more than provide a framework in which we can pose the difficult philosophical questions precisely. These questions mainly concern the constraints that rationality places on the weights $\beta_{t_1}, \ldots, \beta_{t_n}$ that an agent uses.

The most important such question, it seems to me, is whether an agent should assign greater weightings to her other local utility functions the more similar they are to her present local utility function? On the one hand, this would permit her to heavily discount the opinion of a future self that she knows will have shifted significantly from her current self in its moral opinions. For instance, when deciding how to vote, it would release her current left-wing self from the obligation to place much weight on the values of her possible right-wing future self. It also answers another possible existentialist concern: if an agent does not set her weights in this way, then she will end up deferring in large part to value judgments that are not currently her own, thereby rendering her decision inauthentic. Having said that, it is unclear to me whether the existentialist motivation for authenticity militates against deference to one's own future value judgements or merely deference to the value judgements of others or to those of an objective morality. On the other hand, such a weighting can lead to a certain conservatism as well as an unappealing chauvinism or parochialism about one's current values, especially if one thinks that a future local utility function, while very different from one's own, is nonetheless a permissible local utility function to have. If I assign little weight to my future local utility function as a parent because it lies far from my current utility function as a non-parent, and I choose not to adopt on that basis, something has gone wrong. It may be thought that second-order utilities can help solve this problem, but it is worth noting that my second-order utilities can change as well, and we will then need third-order utilities to adjudicate cases in which those might change; and so on. Also, in the case of choosing to become a parent, for instance, it seems that my second-order utilities will change with my first-order utilities: currently, I value valuing friends above family; and, if I adopt, I will value valuing family above friends. Thus, even if we use second-order utilities to assign weights, we'll be left with the same chauvinism about current values that we wished to escape.

Thus, we have seen how deliberative decision theory may be reformulated in order to avoid Paul's challenges, at least formally. However, those challenges do raise profound philosophical questions about the status of decisions we make using that theory. In particular, they raise a number of important questions concerning the extent to which such decision making is compatible with the claims of existentialism.

References

[Paul, 2014] Paul, L. A. (2014). *Transformative Experience*. Oxford University Press, Oxford.