Comments on Katie Steele's "What can we Rationally Value?"

Kenny Easwaran

May 26, 2006

The central point of Katie's paper is the principle of independence in decision theory, and its relation to the Allais paradox. To state independence again, "a rational agent's preference between A and A* should not depend on what happens in circumstances where the two yield identical outcomes." (Joyce, p. 86)

She discusses a few options that have been considered to respond to this situation:

1. Savage suggested a kind of error theory about ordinary decisions in Allais cases, saying that people are just wrong, and we should go by expected utilities.

2. Broome suggested keeping independence, but modifying outcomes by adding an emotional factor or a counterfactual factor. That is, when specifying the outcomes only in terms of money, we have left out the emotional contribution to the utility of the states. Objections to this account point to Savage's "rectangular field assumption" - but Katie has rightly pointed out that this is just an idealization. We don't require every action in this rectangular field to be available, so why should we require them all to make sense?

3. Machina suggested that we give up independence, making decision theory in some sense "non-local" rather than depending just on the states and outcomes. The first objections involve the effect this would have on Dutch books and value additivity (which Machina would also have to deny), while the stronger objections point out how this would affect similar games framed in terms of sequential choices.

In this case, not many people seem to want to stick with Savage in the first response and reject ordinary reasoning in the Allais case as just irrational. But it doesn't seem to me clear why this is - after all, in the example of St. Petersburg (with payoffs suitably modified to account for diminishing marginal utility of money) this is exactly what many people do want to say. Why do we think people's reasoning in the Allais case is so much better?

The second, emotion-considering modification, clearly seems to be the right response to Sen examples, but Katie suggests the emotional response in the Allais case seems to be relatively weak. However, the Allais effect is also pretty weak. Anecdotally, I have actually found we have to be careful in specifying the numbers to get the Allais decisions, so perhaps a relatively weak emotional response is sufficient to generate the actual effect.

Perhaps a better argument is the following modification of the Allais setup - the monetary payouts will be just the same on each of the tickets in each of the four choices. However, in each case, an additional effect will occur - simultaneously with receiving the money (and the information of which payout was received), one will have one's memory modified so as to believe that this particular payout was in fact the payout for all tickets in the game. This should remove the "relief" or "disappointment" factors from each of the payouts, so that the 0 feels no worse in game b than in game d. But intuitively, this does nothing (or almost nothing) to change one's preferences in the situation. Perhaps the violation of rationality involved (ie, knowing that one will lose one's memory of the actual bet involved) may be taken to show that in this example, decision theory just doesn't apply. But it perhaps seems more plausible to suggest that whatever is going on here is exactly what is going on in the regular Allais case, so that counterfactuals, rather than emotions on receiving the payments, are the relevant factor. There is still room for a position on which one's emotions immediately upon picking the game are relevant - but if the memory modification is made soon enough, the effect of these emotions on one's well-being might be able to be made negligible. And if one's emotional feelings about making a bet can be significant in rational decision theory, then decision theory may end up saying nothing beyond "make the choice that feels right". Even for the descriptivist, this would reduce decision theory to observing the emotional responses produced by various bets, rather than thinking about the outcomes of the bets.

Thus, it seems that the emotional response isn't sufficient, so counterfactuals of some sort must be used if one is to support Allais reasoning. However, there are at least two ways of applying them, with differing effects on our notion of independence. The intuition behind independence suggests that all the features of a decision problem that matter should just depend "locally" on the outcomes in particular states. Reading this in a strong way, we can suggest that all that matters are the probabilities and utilities of individual outcomes, that each outcome makes a separate and independent (additive) contribution to the overall utility of a gamble, and that this contribution is linearly proportional to both the utility and the probability of the outcome. With these assumptions, we get expected utility decision theory.

Now, there are at least two ways to respond to the Allais paradox by introducing counterfactuals. The first actually preserves independence in a certain sense - we start with a set of "prima facie utilities" for outcomes, together with a set of probabilities for outcomes, and come up with a set of "actual utilities" for those outcomes, based on how things could have gone otherwise. Then, we make the overall utility proportional to the probability and actual utility, rather than prima facie utility.

The second is more clearly non-local - we apply expected utility theory to the "prima facie utilities" and then add an additional factor based on the particular profile of the gamble. In the first case, we say that the contributions of the outcomes aren't separate and independent, while in the second we say that in addition to the contributions by the individual outcomes, there is one further contribution based on the overall profile of the gamble.

Despite seeming differences, these two approaches are exactly equally powerful - by choosing the transformation function right from "prima facie utilities" to "actual utilities", we can take the amount of any further contribution and add it to each local payout. Similarly, given the modifications on a set of individual states, we can take their expectation, and add this value as the further contribution. Thus, if one trivializes decision theory, then the other does as well, even though they violate the intuitions behind independence in different ways.

Katie points out this risk of trivializing the theory. To spell it out, any (linearly ordered) set of preferences whatsoever can be justified by such an account, by coming up with a gerrymandered function of some sort. For instance, if someone suggests I should rank a coin flip between 1 and -1 as highly as a sure 1,000,000, but follow expected utility otherwise, they can say that the specific outcome of "1, where -1 had an even chance of occurring" actually has utility 2,000,001, while all other counterfactuals are irrelevant. Leaving these functions completely unconstrained thus requires particular empirical work to discover how people actually make decisions, and can give us no normative force whatsoever.

However, carefully used counterfactuals should run no such problems. For instance, we might want to say that whatever the overall effect of counterfactuals, the result should not violate statewise dominance on prima facie utilities. When this sort of restriction is imposed, counterfactuals won't trivialize the account. The specific sort of restriction we need may look more natural either on the local or non-local modification of expected utility theory - so this is going to have to be the way we decide between the two.

Either suggestion will of course give a whole lot more freedom to decision theory, which is still in general a negative for any theory. We want our theories to have as few parameters as possible, so that they have stronger predictive power, rather than just being an ad hoc set of experimental measurements. This concern should be just as strong for the theorist whose picture of decision theory is purely descriptive as for the one with a strong normative take on the matter. Physicists (who are purely in the descriptive business, not the normative one) want their theories to predict fundamental constants, rather than having them as parameters that have to be set, the way a factor for the variance would have to be. Of course, having a principled argument that such a factor must be 1 (or some other nice number) would do much to assuage these worries, but in practice, it seems that most such factors will have to be set based on just how people seem to actually make their decisions - the fact that variance is proportional to the squares of raw magnitudes, while expectation is linear, suggests that there may even be some dependence in just what scale of utility is used.

Right now I have been considering all these decision problems described entirely as sets of outcomes with probabilities. Redescribing them with extensive game trees, as Teddy Seidenfeld and others have pointed out, raises problems of its own. But in fact, Katie makes a good point - any approach of redescribing the problem somehow that has a chance of working seems quite likely to end up being just as complicated and hard for decision makers to keep track of as a modified strategy for these extensive trees.

In conclusion, I'd like to thank Katie for bringing my attention to a particular set of problems in decision theory that I hadn't much considered before. As she says, it seems that whatever solution one proposes will end up having to be quite complicated - there are few savings to be had by describing the solution one way rather than another. The conclusions are still quite ambiguous, but this just means we need further work here.