I

World Bayesianism

1

Utility theory and the Bayesian paradigm

Abstract. A problem for utility theory – that it would have an agent who was compelled to play Russian roulette with one revolver or another, to pay as much to have a six-shooter with four bullets relieved of one bullet before playing with it as he would be willing to pay to have a six-shooter with two bullets emptied – is explained. A less demanding theory that does not have this problem, causal world Bayesianism, is described. This theory would have an agent maximize expected values of worlds in which his actions, complete with their risk dimensions, might take place. Utility theory is located within that theory as valid for agents who satisfy certain formal conditions: It is valid for agents that are in terms of that more general theory indifferent to certain dimensions of risks.

0. Introduction

Utility theory is characterized provisionally in Section 1, after which a problem for this theory – the Zeckhauser–Gibbard problem – is set out. A reaction to this problem that would have one insist on "complete basic alternatives" is then considered, and it is observed that it would make the theory inapplicable to preferences of some reasonable people if certain attitudes to risk are allowed to be reasonable. Next, a general theory based on complete alternatives – "practical worlds" – is formulated, and utility theory is established within it as valid for agents who satisfy certain conditions of risk neutrality. Consideration of Savage- and Raiffa-style arguments against common preferences in our problem is then resisted. Addenda take up implications for game theory and relations between utilities and values. A final "Summing Up" comments on the recent history of the Bayesian idea that would have rational choices and preferences be determined by weighted averages of values of possible outcomes.

1. Utility theory

Let a *lottery* be a distribution of chances to states or lotteries; let a *lottery set* include precisely the members of some set of incompatible basic alternatives: every simple lottery with only basic prizes; every first-order compound lottery with only basic outcomes and simple lotteries as prizes; and, for every n, every nth-order compound lottery with only basic outcomes, simple lotteries, and less than nth-order compound lotteries as prizes. Let a *utility function* be an assignment of numbers to states and lotteries in a lottery set such that the number assigned to a lottery L is its expected value, that is, the weighted average of the numbers assigned to its possible prizes (whether these be lotteries or states) in which the weights employed are the chances L accords directly to these prizes.

Provisionally, we state utility theory, as a theory of ideally rational preferences, thus:

> An agent's preferences for mutually incompatible states are ideally rational only if (i) the agent has pairwise preferences for all members of a lottery set based on these states, and (ii) a utility function represents these preferences.

A somewhat more demanding principle would add: (iii) the agent prefers X from subset S (supposing a free choice among precisely the members of S) only if, for each X' in S other than X, he pairwise prefers X to X' or is pairwise indifferent regarding them. Applications of utility theory generally take for granted the satisfaction of this more demanding principle, but in what follows we are concerned only with the less demanding one.

Luce and Raiffa (1957) explain how relevant pairwise preferences between members of a lottery set might be made explicit, although not all relevant pairwise preferences can be made explicit because, when there is more than one basic alternative state, there are infinitely many lotteries to compare (Luce and Raiffa 1957, p. 15). The idea is that utility functions would represent certain conditional pairwise preferences. An agent who would make one of these explicit supposes that his choice is just between members of a certain pair (i.e., that he can and must choose one of them), and then chooses in imagination. He makes epistemic suppositions – suppositions that operate to shift his epistemic perspective by conditionalization.

For future reference, I note that if an agent has pairwise preferences for all members of a lottery set, and these preferences are represented by a utility function, then they satisfy the following familiar principles addressed to members of this set.

Transitivity of indifference. For any states or lotteries X, Y, and Z, if $X \approx Y$ and $Y \approx Z$ then $X \approx Z$.

Substitution. For any lotteries X and X' and states or lotteries Y and Y', let $Y \approx Y'$. Let X' be like X except that X' includes, instead of the chances X has for Y, like chances for Y'. Then $X \approx X'$.

The intended sense of this substitution principle is clear enough. One route to a more articulated statement would be by way of a formal language for utility theory with reference to which one could speak literally of substitutions of one term (state term or lottery term) for another within a lottery term.

Reduction. For any lotteries X and X', if ultimate chances (perhaps as revealed by computations) for basic alternatives are the same in X and X', then $X \approx X'$.

Here again things could be cleaner in a theory in which terms were distinguished from states and lotteries designated by them. Lottery sets could then be redefined as containing only basic alternatives and all lotteries with only these alternatives as prizes, and the simple/compound distinction would pertain not to lotteries themselves but only to lottery terms for them. And one could say that when one of the lottery terms X' comes from another X by syntactical operations corresponding to the computations alluded to in Reduction, then X is identical with X' and so of course $X \approx X'$. That arrangement would make plainer that Reduction, in contrast with Substitution, is not a substantive condition (cf. Jarrow 1987, p. 100).

2. The problem

What follows elaborates on a problem sketched by Allan Gibbard at a workshop in 1984. The numbers are his, as well as the main lines of the deduction. The problem is a "variant of a problem of Richard Zeckhauser's that Kahneman and Tversky (1979, p. 283) relate" (Jeffrey 1987, p. 227).

2.1. The case for the problem

Suppose that an agent will play Russian roulette either with a revolver whose six chambers contain two bullets, or with one whose six chambers contain four bullets. Suppose further that, whichever gun the agent plays with, he will have a choice whether to play with it as it stands, or to pay certain amounts to have a bullet or bullets removed before playing.

Suppose there is a maximum payment that he would be willing to make in order to have both bullets removed from the first gun – a payment sufficient to move the agent to say, "That much, but no more; if more I would rather take my chances." To make plausible the reasonableness of this assumption, let possible "payments" include not only transfers of assets but also liabilities to penalties, including liabilities to episodes of torture after he has played and if he survives. The payment here consists in the acceptance of a liability to torture. This payment is made whether or not the agent survives and the torture takes place.

Suppose the agent is rich (R) and that, after paying the most he is willing to pay to have both bullets removed from the first gun before playing with it, he would be poor (P). Suppose that he will be dead (D) if and only if he plays and loses. Let him be sure that his chances of dying equal the number of bullets in the gun, divided by 6. And let the agent be exactly indifferent between (i) paying the most he would be willing to pay to have two bullets removed from the first gun, thereby ensuring that he does not die but is rendered poor, and (ii) not paying anything and taking his chances with that gun as loaded:

(1) $$(\sim D \& P) \approx [\tfrac{2}{6}(D \& R), \tfrac{4}{6}(\sim D \& R)].$$

To make plausible the idea that the most he would be willing to pay should strike this balance, we include as possible payments all *chances* of payments. Suppose also that this agent does not care whether he dies rich or poor:

(2) $$(D \& R) \approx (D \& P).$$

Such indifference, while rare (since most people take an interest in their posthumous estates), is not unknown; even if it be ungenerous and inconsiderate, it is not necessarily unreasonable. Suppose now that this agent has pairwise preferences for the four compound states $(\sim D \& R)$, $(\sim D \& P)$, $(D \& R)$, and $(D \& P)$, and for all lotteries based on these states.

Our conditions are so far certainly consistent with the agent's being reasonable. To test the general idea that preferences are reasonable only if they are represented by a utility function, we suppose finally that this agent's preferences are represented by a utility function.

2.2. The problem of this case

The problem lies in what these conditions imply. By Substitution we have, given (1), the indifference

(3) $$[\tfrac{1}{2}(\sim D \& P), \tfrac{1}{2}(D \& P)] \approx [\tfrac{1}{2}[\tfrac{2}{6}(D \& R), \tfrac{4}{6}(\sim D \& R)], \tfrac{1}{2}(D \& P)];$$

by Substitution, given (2), we have the indifference

(4) $$[\tfrac{1}{2}[\tfrac{2}{6}(D\,\&\,R), \tfrac{4}{6}(\sim D\,\&\,R)], \tfrac{1}{2}(D\,\&\,P)] \approx [\tfrac{1}{2}[\tfrac{2}{6}(D\,\&\,R), \tfrac{4}{6}(\sim D\,\&\,R)], \tfrac{1}{2}(D\,\&\,R)]$$

From (3) and (4), by Transitivity we have the indifference

(5) $$[\tfrac{1}{2}(\sim D\,\&\,P), \tfrac{1}{2}(D\,\&\,P)] \approx [\tfrac{1}{2}[\tfrac{2}{6}(D\,\&\,R), \tfrac{4}{6}(\sim D\,\&\,R)], \tfrac{1}{2}(D\,\&\,R)].$$

And from (5), two applications of Reduction and Transitivity yield

(6) $$[\tfrac{3}{6}(\sim D\,\&\,P), \tfrac{3}{6}(D\,\&\,P)] \approx [\tfrac{4}{6}(D\,\&\,R), \tfrac{2}{6}(\sim D\,\&\,R)].$$

This last indifference (6) is the problem, as the following considerations show. The lottery to the left – $[\tfrac{3}{6}(\sim D\,\&\,P), \tfrac{3}{6}(D\,\&\,P)]$ – is the situation faced by the agent who pays to have *one* bullet removed from the *second* gun and who pays precisely his maximum amount to have *both* bullets removed from the *first* gun. Paying that much would make him poor – either not dead and poor, or dead and poor – after playing with the partially unloaded second gun. He would, on paying this amount, have a $\tfrac{3}{6}$ chance of ending up poor and alive and a $\tfrac{3}{6}$ chance of ending up poor and dead: having paid this amount, the agent would be poor, and one of the four bullets would be removed from the gun. Indifference (6) has on the right $[\tfrac{4}{6}(D\,\&\,R), \tfrac{2}{6}(\sim D\,\&\,R)]$, which is the situation faced by the agent who chooses not to pay and to accept a $\tfrac{4}{6}$ chance of dying rich. Indifference (6) says that the agent is indifferent between these situations. In other words,

(6*) The most this agent would be willing to pay to have both bullets removed from the first gun, supposing he were required to play with it, is exactly equal to the most he would be willing to pay to have one bullet removed from the second gun, supposing he were required to play with *it*.

This exact equality follows no matter how much, or how little, is the greatest amount the agent would be willing to pay to have both bullets removed from the first gun, provided only that

(i) his highest price is such that he would as soon pay it as not,
(ii) he is indifferent to the effects on his estate were he to die,
(iii) he has pairwise preferences for all relevant lotteries, and
(iv) *these preferences are represented by a utility function*.

Indifference (6) follows even if the agent is suicidal, so that the most he would be willing to pay is some "negative payment" that would be better described as "the least he would be willing to accept." For clarity in a suicidal case, one might interchange *R* and *P*; accepting a payment to have bullets removed would make a suicidal agent richer, not poorer.

2.3

Should (6*) be true of every agent who satisfies the conditions of the case other than the one being tested? *Must* it be true of any such agent, if he is rational in his preferences? Suppose it is settled that a person must play with the first gun. Couldn't it be reasonable for him to pay more to have both bullets removed from this first gun than he would be willing to pay to have only one bullet removed from the second gun if *it* were the gun with which he was required to play? Surely this variance in acceptable prices could be reasonable for someone who: was not eager to pay his highest price in the case of the first gun (condition (1)); was indifferent to his estate were he to die (condition (2)), so that he does not care whether he dies rich or poor; and had preferences for all relevant lotteries by pairs.

It can indeed seem that a variance in acceptable prices here would be not only possible but *mandatory* for most persons. It can seem that for anyone who had an interest in the outcome, it could not be reasonable to pay exactly as much to improve his chances for life from $2/6$ to $3/6$ as to improve his chances from $4/6$ to a certainty. But this strong conclusion cannot be maintained. Consider a person who wants to live. He might be willing without error or unreason to pay as much for the smaller improvement from $2/6$ to $3/6$ in chances for life, just because the smaller improvement in his chances for life would take place in circumstances where his chances of dying were in any event substantial, and he believes that "You can't take it with you." (Recall that our agent is of a can't-take-it-with-you turn of mind: he does not care whether he dies rich or dies poor.) It may be asked, "How *could* he be willing to pay as much to have just one bullet removed from the four-bullet gun as he would be willing to pay to have both removed from the two-bullet gun?"; one possible answer is that the money paid (or the risk incurred of torture) in the first case could be worth less to him (cf. Kahneman and Tversky 1979, p. 283). It evidently can be reasonable to pay as much to have just one bullet removed from the second gun as one would pay to have both removed from the first gun. (I owe recognition of this point to Włodek Rabinowicz.)

Although we can't say that the price for having two bullets removed from the first gun would, for *any* reasonable agent who wanted to live, be too high a price to pay for having just one bullet removed from the second gun, this price would surely be too high for *some* reasonable agents.

I believe that many reasonable agents who wanted to live would, contrary to indifference (6), prefer not to pay this price, and would persist in this preference even after thoughtful reflection, and that the reverse would prove true of many reasonable agents who wanted to die.

An argument for indifference (6) goes through, we may note, even for an agent who does not care whether he lives or dies. The most such an agent would be willing to pay to have the first gun emptied would be nothing, so that $P = R$ and thus (i) $(\sim D \& R) = (\sim D \& P)$. Furthermore, since he does not care whether he lives or dies, presumably (ii) $(\sim D \& P) \approx (D \& P)$. Indifference (6) is an easy consequence of (i) and (ii) and principles of utility theory. However, in this odd case in which agents satisfy our initial conditions, (6) says what one expects, and is not a problem for utility theory. The problem posed by indifference (6) is that it seems *not* to say what must be true of every reasonable agent, including of course every thoughtful and reflective agent who does care whether he lives or dies.

3. A reaction: Insist on complete basic alternatives

3.1

The problem, one may feel, is not with utility theory as a theory of ideally rational preferences, but only with that theory as provisionally stated; or, equivalently, with the over-hasty application of the theory in our case. Utility theory is perhaps properly applicable only to preferences for members of lottery sets founded on alternatives that are, relative to all relevant pairwise comparisons, *complete* with respect to things of interest to the agent. The suggestion with regard to our problem could be that its threat is only a prima facie one, since the four compound states $(\sim D \& R)$, $(\sim D \& P)$, $(D \& R)$, and $(D \& P)$ can be expected to be not relevantly complete nor the "real alternatives" in the situation (Luce and Raiffa 1957, p. 28). For example, it certainly seems that for many agents it would *not* be all the same whether they lived poor, $(\sim D \& P)$, when they had a choice between this for sure and taking the chance $[\frac{2}{6}(D \& R), \frac{4}{6}(\sim D \& R)]$, or lived poor after taking the chance $[\frac{3}{6}(\sim D \& P), \frac{3}{6}(D \& P)]$ when they could instead have taken the chance $[\frac{4}{6}(D \& R), \frac{2}{6}(\sim D \& R)]$. It seems that reasonable agents' attitudes toward the four basic outcomes of the problem can be sensitive to shapes of lotteries from which these alternatives might issue, and to shapes of alternative lotteries. But then, according to the present idea, utility theory cannot be tested by applications to preferences of such agents for lotteries based on our four relatively simple compound states; by hypothesis, these states would not be complete for such agents.

3.2

Perhaps utility theory should be addressed only to lottery sets founded on complete basic alternatives that are (in the way just now indicated) complete, and its provisional statement amended thus:

> An agent's preferences for any mutually incompatible states *that are complete with respect to all pairs of members of a lottery set based on them* are ideally rational only if he has pairwise preferences for all members of a lottery set based on these states, and these pairwise preferences are represented by a utility function.

If this is right then the problem in our case is with its over-hasty application of utility theory, though by this I mean not merely that the application discussed would be improper for some persons but that, for at least some reasonable persons, utility theory would be *quite inapplicable.* The theory as amended is applicable only if all lotteries founded on certain basic alternatives are *internally coherent,* and are by pairs *coherently comparable.* But it seems that for some rational agents and situations there may be no set of basic alternatives satisfying these logical conditions and satisfying also the condition that its members are complete (or fully specific) with respect to all things of interest to the agent in pairwise preferences for lotteries based on them.

Suppose, for example, that it matters to an agent whether or not A obtains as an outcome of a lottery in which A and B have chances of ½. Let A' be the version of A in which it is an outcome of such a lottery. Then neither [⅓(A'), ⅔(B)] nor (assuming that C does not entail B) [½(A'), ½(C)] is internally coherent. In contrast, both [½(A'), ½(B)] and (understanding B' similarly to A') [½(A'), ½(B')] are internally coherent. Suppose next that it matters to an agent whether or not basic alternative A results in a situation in which he can choose either A or a lottery L in which A is not a possible outcome. Let A'' be A in such a situation – that is, A when lottery L could have been chosen (i.e., "A instead of L"). Then, for $L' \neq L$, A'' and L' are for this agent not coherently comparable, for it is not logically possible for him to have a choice between just A'' and L' (as A'' is open to choice only when L is), and so he cannot (for purposes of a shift of his epistemic perspective by conditionalization) suppose that his choice is between just A' and L' and then choose in imagination. Objection: "But these are not reasonable attitudes; it should not matter by what means one obtains A, or what one's alternatives are." Response: "Oh?"

Conditions for lottery sets of comparable members can be expected to clash with the requirement that basic alternatives be complete, when an

Cambridge University Press
978-0-521-03898-0 - Taking Chances: Essays on Rational Choice
Jordan Howard Sobel
Excerpt
More information

agent's preferences "reflect intrinsically comparative views of payoffs" (Jeffrey 1987, p. 225). This clash will occur when preferences do that in a certain manner: It will happen when an agent's preferences are informed by certain attitudes toward the shapes of lotteries and alternatives – certain attitudes that are not derivative of how he thinks he would feel afterwards (e.g., regretful, chagrined, delighted) if he were to win or lose. (Recall that, in the problem, the agent can be supposed to know that he will not feel regret if he loses – that he will not feel anything.)

It may be useful to recall a distinction. In one case, an agent realizes that if he were to run a risk and lose then he would regret running the risk and be bothered, but *not* because he would think it had been a mistake (bothersome feelings of regret aside). Here we have a "minimalist" interpretation of the idea of a preference's "reflecting an intrinsically comparative view of payoffs" (Jeffrey 1987, pp. 225–6). In a second case, an agent realizes that if he were to run some risk and lose then he would regret having run it and also be bothered because he would realize that it *had* been a mistake to run it (bothersome feelings of regret again aside). Suppose an agent thought that taking two aspirins would ensure that he would in no case experience regret. He might reasonably take them and run the risk in the first case (depending on his view of the merits of the risk, threats of regret aside), but perhaps not in the second. His aversion in the first case is based partly on matters extrinsic to the risk – on a probable, and conceivably blockable, consequence of running it. In the second case his aversion can, for all that has been said, be entirely intrinsic and based on the risk's nature, all merely possible consequences of running it quite aside (cf. Sobel 1988c, p. 542, n. 6). Whether or not it is irrational to be prone to regrets of the first kind, they do not make theoretical problems; extrinsic psychological effects can enter into payoffs of all possible lotteries, and are consistent with all possible alternative pairings of lotteries. In contrast, regrets of the second kind reflect intrinsically comparative views of payoffs interpreted in nonminimalist ways, views that can create problems for the applicability of utility theory.

3.3

It has been said that when applying utility theory to a case "it may be *necessary* to use a richer set of basic alternatives in order for [the theory] to be . . . valid" (Luce and Raiffa 1957, p. 29, emphasis added). The problem I am pressing is that it may not always be *possible* to enrich an initial set of basic alternatives in all relevant ways. Making basic alternatives complete with respect to things of practical interest to a given rational agent, can, depending on what things interest him, be inconsistent with