Taking unconsidered preferences seriously*

Robert Sugden

School of Economics

University of East Anglia

Norwich NR4 7TJ

United Kingdom

19 September 2005

* An early version of this paper was presented at a conference on philosophical aspects of social choice and welfare, held at the University of Caen. I thank participants at this conference, particularly Kaushik Basu, for valuable comments. I also thank Serena Olsaretti for her suggestions as editor.

In normative economic analysis, it is conventional to treat each person's preferences as that person's own standard of value, and as the standard by which the effects of public policies on that person should be valued. The proposal that preferences should be treated in this way is usually qualified by two apparently natural conditions – that preferences are internally coherent, and that they reflect the considered judgements of the person concerned. However, there is now a great deal of evidence suggesting that, in many economic environments, preferences of the required kind simply do not exist. It seems that the preferences that govern people's actual behaviour are often incoherent and unstable. This prompts the following question: Is there a defensible form of normative economics which respects each individual's *actual* preferences, whatever form they take? I shall try to show that there is.

1. A contractarian approach to normative analysis

This paper is premised on a contractarian understanding of normative analysis. On this understanding, the object of normative analysis is not to arrive at an all-things-considered judgement about what is valuable for society as a whole, but to look for proposals that each individual can value from his or her own point of view. To follow this approach is to treat social value as *subjective* and *distributed.* To say that social value is subjective is to say that it is not a property of the world, but a perception or attitude on the part of the valuer. To say that social value is distributed is to say that it is not a single measure, expressing a synoptic judgement about what is valuable; it is simply an array of the separate value judgements of the individuals who comprise society. There is, then, no distinction between a person's own standard of value and the standard by which effects on that person are valued.[1]

A contractarian theory needs a way of representing the standard of value of each individual in a society of many individuals. In this context, a broad-brush approach is unavoidable. Clearly, we cannot expect to describe the actual value judgements of particular people: we must be content with a theoretically tractable representation of the standard of value of a typical individual. However, the subjectivity of the contractarian approach requires that this representation is flexible: it should impose as few restrictions as possible on the substantive content of any individual's values.

A contractarian theory is addressed to individuals as reflective citizens. The aim is to recommend proposals for collective action by showing that they promote the values of each citizen, considered separately. There is an implicit presumption that these citizens are willing to engage in the kind of reasoned argument that the theory exemplifies, and are capable of recognising the validity of what has been shown. Thus, the values that are being promoted should be understood as the considered judgements of the relevant individuals, and not merely as pre-reflective hunches. It seems natural, then, to represent each individual's standard of value as having some degree of internal coherence.

However, it is fundamental to the contractarian approach that we (the theorists) take moral psychology as it is, not as we would wish it to be. We must not be tempted to suppose that our own favourite moral positions are the only ones that are capable of being sincerely endorsed by people who have engaged in serious reflection. Nor, since contractarianism is premised on a subjective understanding of value, should we require considered judgements to be supported by reasons which appeal to a conception of moral truth, or which are shared by an imagined community of right-thinking moral agents. Rather, we should think of a person's considered judgements as codifications of his perceptions of value, recognising that those perceptions are products of human psychology and social learning, not insights into a moral universe. When, later in this paper, I speak about finding coherent *formulations* of particular moral intuitions, it is this kind of codification that I have in mind.

How, then, should standards of value be represented in a contractarian theory? The most obvious solution, one might think, is to follow the conventions of normative economics and identify each person's standard of value with his considered preferences. In the next two sections of this paper, I examine this proposal and show how it breaks down when individuals lack stable and coherent preferences. Then I suggest an alternative standard of value, based on considered judgements about opportunity rather than on considered preferences.

2. Considered preference as a standard of value

If a person's preferences are to serve as a standard of value, it seems unexceptionable to require that those preferences should have whatever properties of coherence are intrinsic to the concept of value. In normative economics, these properties are usually presented in purely formal terms. Exactly what is required depends on the nature of the universe X of objects of choice over which preferences are defined. If, for example, X is defined as a finite set of discrete objects, it is normal to require that the binary relation [≥] ('is weakly preferred to') is complete (i.e. for all x, y in X: x [≥] y or y [≥] x) and transitive (i.e. for all x, y, z in X: [x [≥] y and y [≥] z] implies x [≥] z). Additional properties are imposed if X is defined as the set of n-dimensional vectors of non-negative real numbers (interpreted as 'bundles' of n consumption goods), or if it is defined as the set of all probability mixes of some set of 'consequences'.

To illustrate the justification for these restrictions, take the case of transitivity. One might argue that transitivity is intrinsic to the concept of value: if x is at least as valuable as y, and y is at least as valuable as z, then, as a matter of logic, x is at least as valuable as z.[2] It is much less plausible to claim that transitivity is intrinsic to the concept of preference. Whether a person's preferences are transitive or not seems to be an empirical matter. Non-transitive preferences might be a symptom of irrationality, but there is no logical contradiction in supposing their existence. So, if we are to use preferences as a standard of value, we have to be sure that they satisfy transitivity.

It is perhaps misleading to think that what is required here is simply a formal restriction on preferences. We cannot make sense of a formal restriction like transitivity without first postulating a particular universe of objects of choice. The economic theory of choice has content only because, at any given time, there are established conventions about how objects of choice are defined. These conventions reflect the role that preferences play in normative analysis. Objects of choice have to be specified in such a way that it is credible to treat them as carriers of value. For example, suppose that Tom, if given a straight choice between an apple and an orange, would prefer to take the apple. Given a straight choice between the orange and a pear, he would prefer to take the orange. But, given a straight choice between the pear and the apple, he would prefer to take the pear. If we follow the conventions of economics and define X = {apple, orange, pear, ... }, we have a violation of transitivity. But what if we define X so that it includes options such as 'taking the apple when there is a straight choice between an apple and an orange'? If we do this, Tom's preferences will come out as transitive. However, unless some explanation can be given for why, from Tom's point of view, the value of a particular fruit differs according to the set from which it is chosen, the claim that these preferences express a standard of value seems unpersuasive.[3] It seems that, as a necessary condition for a person's preferences to be treated as a standard of value, they must satisfy appropriate conditions of coherence when defined over a *normatively credible* universe of objects of choice.

A similar argument can be made about the stability or volatility of preferences. The economic theory of choice has content only because a person's preferences can be assumed to be reasonably stable over time. If they did not have that kind of stability, their credibility as carriers of value would be open to question. Suppose that at 10:00 on some day, Jane is offered a choice between pasta and chicken for a meal to be eaten in the evening, and she expresses a strong preference for pasta. At 10:30, she is given the opportunity to reconsider her choice, and she expresses a strong preference for chicken. On one interpretation of these observations, they disconfirm the hypothesis that Jane's choices are governed by coherent preferences. Alternatively, we might say that Jane has one preference ordering at 10:00 and a different one at 10:30. But, unless some explanation can be given for why, from Jane's point of view, the value of chicken and pasta as evening meals differs according to the time in the morning from which they are viewed, it seems hard to claim that each of these conflicting preferences expresses her values. Some degree of stability of preference seems to be necessary before preferences can be treated as a standard of value.

This thought provides the justification for a restriction proposed by David Gauthier (1986, pp. 26-38) in formulating his contractarian theory of 'morals by agreement': that preferences must be *stable under experience and reflection*. As Gauthier emphasises, this formulation is consistent with a subjective and distributed conception of value: there is no appeal to any notion of value external to the individual's own perceptions. Nor is there any appeal to counterfactual hypotheses about what an individual would prefer, were she to have true knowledge of the world, as in 'informed desire' theories of value. Gauthier's restriction captures a qualification which, I think, would be accepted by most proponents of the principle that preference is a standard of value.

These qualifications can be brought together as a concept of *considered preference*. I shall say that a person's preferences are considered if they satisfy conventional properties of coherence when defined over a normatively credible universe of objects of choice, and if they are stable under experience and reflection. If considered preferences exist, they can be used as the basis for a theory of subjective and distributed social value: we can assert that each person's standard of value is given by her considered preferences. But do they exist?

3. Do considered preferences exist?

At the conceptual level, there is no guarantee that considered preferences exist at all. One might reasonably assert that, if an agent is capable of autonomous choice, she must have preferences of some sort. That is, she must have attitudes towards the options between which she has to choose, and those attitudes will be revealed in some way in the choices that she actually makes. Those attitudes, we might say, are the agent's preferences. However, there is no conceptual guarantee that those preferences satisfy the conventional coherence requirements when defined over some normatively credible universe of objects of choice. Nor is there any guarantee that they are stable under experience and reflection. Whether preferences have these properties is essentially an empirical question. The viability of the considered preference account of value depends on the truth of the hypothesis that, for most people and for most economically significant choices, considered preferences exist.

For the last twenty-five years, the investigation of this hypothesis has been a central research programme of what has come to be known as *behavioural economics* – that branch of economics which draws on the theoretical ideas and experimental methods of psychology to investigate the actual behaviour of economic agents. As a result of this research programme, we now know that economic agents often do *not* act on considered preferences.

Many highly predictable patterns of behaviour have been discovered which contravene received assumptions about preferences. For example, individuals' decisions in controlled experimental environments show a surprisingly high degree of unexplained stochastic variation: if the same person faces exactly the same binary choice problem twice within a few minutes, the probability that she will choose differently in the two cases is of the order of 25 per cent.[4] If preferences are defined over the normatively relevant domains that economic theory has traditionally used, some of the most fundamental principles of preference coherence – including transitivity, dynamic consistency, and the principle that preferences over lotteries should respect stochastic dominance – are systematically violated. In many cases, preferences are highly sensitive to what appear to be normatively irrelevant matters of presentation. Although many of these violations of standard theory were first observed in laboratory environments, they have been found to occur in 'real' markets too.[5]

Just how serious these 'anomalies' are remains a matter of dispute. While some economists conclude that the received theory is fundamentally flawed, others claim that most anomalies result from errors; given sufficient experience and feedback, it is said, individuals correct those errors and 'discover' their underlying preferences.[6] To propose this *discovered preference* hypothesis is, in effect, to propose that considered preferences exist, even if they are not always revealed in choice behaviour (whether in laboratory experiments or in real markets). The evidence on this issue is mixed; while some anomalies do seem to become less frequent as individuals gain experience, others seem much more robust.[7] For the purposes of this paper, it is not necessary to adjudicate these disputes. It is sufficient to recognise that the existence of considered preferences is not a self-evident truth or a well-established empirical fact: it is merely a contested hypothesis. It is surely worth asking what kind of normative economics would be possible if that hypothesis had to be rejected.

Readers who are not familiar with the literature of behavioural economics may find it surprising that apparently unexceptionable principles of rationality are regularly violated. To understand this, one must recognise that, in many cases, the supposed irrationality of these anomalous patterns of behaviour is not visible in any single decision. As an illustration of this general point, consider the phenomenon of 'coherent arbitrariness' reported by Dan Ariely, George Loewenstein and Drazen Prelec (2003).

In one of Ariely et al's experiments, subjects were exposed to an annoying sound and then asked whether, hypothetically, they would repeat the same experience in return for a specified payment. For each person, this hypothetical payment was constructed from the first three digits of her social security number (so that, for example, a person whose number began with 356 was asked to consider a payment of $3.56); subjects knew that this procedure was being used. Following this, subjects were asked to state the minimum amounts of money they would accept as payment for listening to the sound for various lengths of time. 'Prices' were then drawn at random from a distribution, and subjects who had indicated their willingness to listen to the noise at the relevant price were required to do so and were paid the price. This procedure generated a surprising result: subjects whose social security numbers were above the median asked for 60 per cent more payment than those whose numbers were below the median. The implication is that subjects' expressed preferences between money and noise were strongly influenced by the sum of money referred to in the hypothetical payment question, even though it must have been obvious that this carried no information relevant to the task at hand, and even though subjects knew exactly what the experience of noise would be like.

It may seem absurd for a person's preferences to be determined by her social security number but, if we look at the behaviour of any individual subject in this experiment, we see nothing obviously irrational. Consider two typical (but imaginary) subjects. Alan, whose social security number happens to begin with 204, states a willingness-to-accept valuation of

$3.55. Betty, whose number begins with 835, states a valuation of $5.76. There is nothing irrational about either valuation, considered in isolation. Probably, neither person is conscious that his or her valuation has been influenced by the irrelevant cue. Viewed in the perspective of rational choice theory, what is wrong with preferences like these is that they are unstable: if examined carefully, they can be found to vary according to factors which seem to have no rational significance.

One might think – indeed, I do think – that such preferences are not *ir*rational at all, but merely *a*rational. If social security numbers have no rational significance for valuations, rationality cannot prescribe valuations which depend on social security numbers. But what if rationality does not prescribe *any* unique valuation? If it is compatible with rationality to value the noise *either* at $3.55 *or* at $5.76, why is it contrary to rationality to report the first valuation if one has one social security number and the second if one has another?   But I do not want to get sidetracked into the metaphysics of rationality. What matters is that violations of the principles of rational choice theory need not reveal themselves as obvious pathologies.

Still, the fact remains that the preferences that govern choices in Ariely et al's experiment are not stable under experience and reflection. It may well be that, in cases like this, considered preferences simply do not exist. Given one arbitrary stimulus, people are conscious of one preference; given another, equally arbitrary stimulus, they are conscious of a different one. There may be nothing more to be discovered than this. Clearly, these findings are bad news for the programme of using considered preference as the standard of value. But should we conclude that when people act on *un*considered preferences, it doesn't matter to them whether those preferences are satisfied? I suggest not.

Consider an example from everyday life. It is a hot day. I go into a shop to buy a newspaper. I pass a cabinet displaying chilled cans of sweet drinks but not, unfortunately, chilled unsweetened drinks. The idea of a cold drink is attractive to me (I'm hot and thirsty), but the idea of sweetness is aversive (normally, I dislike sweet drinks). I am conscious of opposing motivational pulls. I do not perceive either pull as obviously stronger than the other, but nor do I perceive them as exactly balanced, as I would if I were choosing between two apparently identical cans of the same brand of drink. I'm just not sure which of the two options ('drink' and 'not drink') to go for. In the end, I go for the drink. Suppose the truth (known to experts in marketing, but not to me) is that people like me are susceptible to the placement of products in display cabinets; had the drinks been on a lower shelf, I would still have seen them, but would not have chosen to buy one. In buying the drink, then, I am not acting on a considered preference. Even though I don't know the marketing theory, I may be conscious that my preferences are unstable: having found it hard to choose between the two options, I may realise that in other, apparently similar circumstances, I might have chosen differently. So is there any value to me in my being able to satisfy my unconsidered preference for the drink?

Speaking for myself, I *do* value such opportunities. To the extent that the economy is structured so that I can satisfy my desires, as I experience them, and however arbitrary and unstable they may be, that is for me something to value. And although the preferences that are being satisfied may be unconsidered, *my valuing the opportunity to satisfy them* is considered: I find that value judgement to be stable under experience and reflection. That judgement may not be shared by everyone, but it is not merely a personal idiosyncrasy. It is the core idea to which generations of economists have appealed when they have argued that competitive markets implement the value of consumer sovereignty, and which is reflected in the business maxim that the customer is always right. Whatever consumers want and are willing to pay for, whether their reasons for wanting it are good, bad or non-existent, producers will find it in their interests to supply. For those of us who value consumer sovereignty, that is one of the great virtues of the market system.

A critic might ask how it can matter to me whether a preference of mine is satisfied, unless I can provide some reason for that preference. Surely (the critic says) what matters to me is that I get what has value, not what I happen to feel a sense of desire for. Preferences that are not stable under experience and reflection are arbitrary mental states; they do not have supporting reasons. Perhaps some higher-order value, such as autonomy, is served by my being free to choose my own actions even in the absence of reasons for choice; but then it is the higher-order value that provides the reason for valuing consumer sovereignty. Merely wanting something (the critic concludes) is not a reason for getting it.

My response is that the critic's demand for reasons is misplaced. As I argued in Section 1, a contractarian approach cannot demand that considered value judgements are justifiable by appeals to reason. When considering a claim about what people value, the contractarian theorist should ask only whether the claim is credible as moral psychology, and whether the alleged standard of value can be formulated in a coherent way. I see nothing incredible in the suggestion that, for many people, being able to satisfy their own desires, as and when they experience them, is something that they perceive as valuable – without their ever imagining that this sense of value needs further rationalisation. Whether this perception can be given a coherent formulation is the subject of this paper.

To avoid misunderstanding, I must make clear that I am not asserting that, for most people, consumer sovereignty always trumps other considerations. One can value consumer sovereignty in general and still recognise certain specific restrictions on one's freedom of choice as being in one's long-term interest. For example, suppose I enjoy moderate consumption of alcohol, but know that this impairs my judgement. Then I might support laws which restrict people's capacity to take decisions with serious long-term consequences while under the influence of alcohol. Everyone will have his or her favourite examples of justified forms of self-constraint. The essential feature of such examples is that, when forming considered judgements about what one values, one disavows certain preferences that one might sometimes wish to act on. All I want to claim is that, for those of us who value consumer sovereignty, such cases are the exception rather than the rule. In the normal run of events, there will be many cases in which one can expect one's preferences to be unstable and yet have no wish to disavow them.

My aim is to consider how this idea might be formulated and, in so doing, to test its conceptual coherence. I do this by taking a case in which a person's preferences are unstable, but in which (for me at least) the intuition in favour of consumer sovereignty remains strong. In this paper, I do not consider the exceptional cases in which there is a considered judgement in favour of self-constraint.[8]

4. Continuing identity: the problem

The focus of my analysis is a stylised model of a simple decision problem for a representative economic agent. I call this model *Joe's problem*, since Joe is the name of my agent. It is constructed to exhibit a transparent violation of the hypothesis that preferences are stable under experience and reflection. My objective is to find a coherent formulation of the idea that the agent values opportunities to act on his unconsidered preferences, even though he may use those opportunities in an apparently inconsistent way.

I begin with the version of the model in which Joe has the greatest freedom of action. There are three time periods. The model concerns Joe's choices with respect to a ticket in a lottery which will be drawn in period 3. The ticket offers a 1 in 100 chance of winning a prize of £1,000. Initially Joe has not entered the lottery, and his only opportunity to do so occurs in period 1; in that period, he has the opportunity to buy the ticket at a cost of £11. If he buys, then in period 2 he has the opportunity to sell back the ticket, but receiving only £9 in exchange. There is no particular significance to these precise probabilities and amounts of money. What matters is that the buying price is higher than the selling price, that Joe might reasonably value the ticket at more than the buying price, but might equally reasonably value it at less than the selling price; and that if he acts on the first valuation in period 1 and on the second in period 2, he will incur an unambiguous loss.

The next feature of the model represents a psychological regularity for which there is solid evidence: other things being equal, people tend to be more willing to take risks in unhappy moods than in happy ones.[9] For Joe, there are just two possible moods, 'happy' and 'unhappy'. In each period, he is one or other of these moods. Which mood he is in is determined by a random process, independently for each period, such that each mood is experienced with probability 0.5. Joe cannot predict his mood in advance. In the unhappy mood, his attitude to risk is such that he perceives the lottery ticket as more desirable than the certainty of £11. In the happy mood, his attitude to risk is such that he perceives the lottery as less desirable than the certainty of £9. Notice that Joe's happiness or unhappiness, as experienced when he makes decisions about buying or selling the ticket (that is, in periods 1 and 2), is independent of his mood at the time the lottery is drawn (period 3). Thus, his different attitudes to risk – his ex ante perceptions about the merits of risky decisions – are independent of the ex post experiences to which those decisions can lead.

In decision theory, it is standard to represent sequential decision problems of this kind as *decision trees*. Each point in the problem at which a choice is required is represented by a *choice node*, conventionally drawn as a square. Each point at which chance intervenes is represented by a *chance node*, drawn as a circle. If the agent's preferences are different at different times or in different contingencies, this is represented by dividing the agent into two or more *selves*. Each self is then modelled as if it was a distinct agent with its own preferences. In this way, the preferences of each self can be held constant throughout the tree.

Applying this strategy to Joe's problem, we can define two selves, an unhappy and risk-loving *Joe_1* and a happy and risk-averse *Joe_2*. To avoid unnecessary complications, I assume that Joe's preferences, for any given mood, satisfy the axioms of expected utility theory. Thus, Joe's preferences can be represented by two utility functions, one for each mood. To keep the decision tree as simple as possible, I start the analysis after Joe's mood for period 1 has been determined, and I stipulate that this mood is unhappy. I end the analysis at the end of period 2, before the lottery has been drawn. Payoffs are represented in units of expected utility.

With the model specified in this way, there are three possible outcomes: *either* Joe doesn't buy the ticket, *or* he buys it in period 1 and holds it through period 2, *or* he buys in period 1 and sells in period 2. I stipulate that, for the unhappy Joe$_1$, the expected utilities of these outcomes are respectively 0, –1 and 2, while for the happy Joe$_2$, they are 0, –1 and –2. This gives us the tree T_1, shown in Figure 1. For each outcome, the expected utility for Joe$_1$ is shown, followed by that for Joe$_2$.

What will Joe do? The standard decision-theoretic analysis uses the *folding-back* (or *backward induction*) algorithm, working out what happens at the most remote nodes of a tree (on the supposition that they are reached) and then working back towards the initial node. So, suppose that Joe$_1$'s second decision node is reached. Risk-loving Joe$_1$ will then choose *hold* (with an expected utility of 2) rather than *sell* (–1). Next, suppose that Joe$_2$'s decision node is reached. Risk-averse Joe$_2$ will choose *sell* (–1) rather than *hold* (–2). Finally, consider Joe$_1$ at the initial node, and assume that he can replicate the analysis we have just been through. So Joe$_1$ knows that if he chooses *buy*, there is a 0.5 probability that Joe$_2$ will *sell*, giving Joe$_1$ a payoff of –1, and a 0.5 probability that Joe$_1$ will *hold*, giving a payoff of 2. From this, it follows that the expected payoff to Joe$_1$ from *buy* is 0.5. Since this is greater than the payoff of *don't buy*, Joe$_1$ will *buy*. So the answer to the original question is that in period 1, Joe will *buy*; in period 2, he will *sell* if he is happy but *hold* if he is unhappy.[10]

Notice that, with probability 0.5, Joe$_1$'s choice in period 1 is undone by Joe$_2$'s choice in period 2. This combination of actions leads to an unambiguous loss. (More formally, the outcome of *buy* followed by *sell* is strictly worse than the outcome of *don't buy*, whether evaluated in terms of Joe$_1$'s preferences or in terms of Joe$_2$'s.) Does this imply that it would be better for Joe, as a continuing person, if the choices of one or both of his selves were constrained? One possibility is to remove the *sell* option in period 2, so that Joe$_1$'s period 1 decision cannot be undone. This gives the contracted tree T_2, shown in Figure 2. An alternative possibility is to remove the *buy* option in period 1, giving the contracted tree T_3. I can now reveal where this story is leading. I want to pose the following question: *Which of the three trees is most valuable to Joe*?

It seems natural to say that the principle of consumer sovereignty favours the unconstrained tree T_1. If Joe is to be a sovereign consumer, he should be free to buy and sell as he chooses. T_2 removes an options to sell, while T_3 removes an option to buy. Each of these constraints prevents Joe from choosing to do one thing in period 1 and then choosing to undo it in period 2. Of course, he would incur a cost in changing his mind in this way. Still, if we are to treat him as a sovereign consumer, we must surely allow him the privilege of changing his mind, provided he is willing to pay the price of doing so. The problem is to find a way of saying that this form of sovereignty is valuable *to Joe* – of saying that, for Joe, T_1 is the most valuable of the three trees. That is the problem I shall now try to solve.

In thinking about this issue, we immediately confront the difficulty that we need to define the continuing Joe for whom trees can have value. On what I take to be the most conventional understanding of multiple-selves models in decision theory, there is no such entity. Of course, there is Joe the continuing human being; but, since preference is the standard of value, an entity can be a locus of value only if it has stable preferences. The only loci of value are Joe$_1$ and Joe$_2$. Taking the viewpoint of a given self, we can rank the trees in terms of their outcomes, as evaluated by that self's preferences. Following this approach, we can say that, for Joe$_1$, the most valuable tree is T_2. (This gives Joe$_1$ an expected payoff of 2, compared with 0.5 from T_1 and zero from T_3.) For Joe$_2$, the most valuable tree is T_3. (This gives Joe$_2$ an expected payoff of zero, compared with –1.5 from T_1 and –2 from T_2.) Each self values the imposition of constraints on the other: Joe$_1$ wants to prevent Joe$_2$ from selling, while Joe$_2$ wants to prevent Joe$_1$ from buying. There seems to be no other Joe for whom the absence of constraints can be valuable: there seems to be no continuing person for whom T_1 is the most valuable tree.

One way of trying to define a continuing person is to appeal to *metapreferences* – that is, preferences over preferences. The idea is that there is a continuing Joe who in some sense prefers, or identifies with, the preferences of one of his selves even while he is acting on the preferences of the other. In recognition of the fact that this continuing Joe is the locus of metapreferences, let us call him Joe^M. We might suppose that Joe^M identifies with the preferences of the happy and risk-averse Joe$_2$, and treats Joe$_1$'s inclination to act contrary to those preferences as weakness of will. On this account, the most valuable tree for Joe^M is T_3. Conversely, if Joe^M identifies with Joe$_1$, the most valuable tree is T_2. Notice that, whatever we assume about Joe^M's metapreferences, T_1 cannot be the most valuable tree. To put this conclusion another way, the special feature of T_1 is that it allows (and, given the actual preferences of Joe$_1$ and Joe$_2$, can induce) the sequence of actions in which *buy* is followed by *sell*; but Joe$^{\mathrm{M}}$ cannot identify with this sequence. Whatever his metapreferences, he must attribute one of the two actions in that sequence to a self whose preferences he disavows.[11]

A different way of conceiving of a continuing Joe is as the *set* of selves {Joe$_1$, Joe$_2$}, and to treat the welfare of this entity (let us call it Joe$^{\mathrm{S}}$) as a weighted average of the utilities of its component selves, on the analogy of a utilitarian social welfare function. Formally, let U^{S} be the welfare of Joe$^{\mathrm{S}}$ and let u_1 and u_2 be the utilities of Joe$_1$ and Joe$_2$; then $U^{\mathrm{S}} = \alpha u_1 + (1 - \alpha)u_2$, where α (the weight given to Joe$_1$) is in the interval $0 < \alpha < 1$. It is easy to work out that if $\alpha > 0.5$, the best tree for Joe$^{\mathrm{S}}$ is T_2; if $\alpha < 0.5$, the best tree is T_3; if $\alpha = 0.5$, T_2 and T_3 are jointly best. There is no value of α at which T_1 is best. That is because, whatever relative weights Joe$^{\mathrm{S}}$ gives to the utilities of his two selves, the outcome of the sequence in which *buy* is followed by *sell* is inferior to that of *don't buy*. For Joe$^{\mathrm{S}}$, the fact that T_1 can induce this sequence is a reason for rejecting it in favour of one of the other trees.

Summing up the argument so far, conventional decision-theoretic analysis seems to be unable to represent the idea of a continuing Joe for whom T_1 is the most valuable tree. But, for me, the intuition persists that, if Joe values his sovereignty as a consumer, he can see T_1 as the most valuable tree. The problem remains: we need to find a way of representing that intuition as a coherent theoretical principle.

5. Continuing identity: a solution

I suggest that we need a radically different conception of the continuing person. We should think of the continuing Joe – let us call him Joe* – as the *composition* of the selves which perform the various parts of whatever sequence of actions is in fact performed. For theorists who insist on modelling identity in terms of some kind of preference relation, this idea may seem strange. But, viewed from outside the framework of decision theory, it seems a very natural way of thinking of identity. The continuing Joe* is just whatever Joe the human being is over time. What the continuing Joe* does is just whatever Joe does over time; what the continuing Joe* values is just whatever Joe values over time. In this perspective, it becomes clear how the continuing person can value the absence of constraints on his present and future actions.

Suppose the decision problem is T_1. In period 1, Joe$_1$ wants to choose, and does choose, *buy*. Since Joe* *is* Joe$_1$ at this moment, it is also true that Joe* wants to choose, and does choose, *buy*. In period 2, let us suppose, chance selects the happy mood. Then Joe$_2$ is the agent, and he wants to choose, and does choose, *sell*. Since Joe* *is* Joe$_2$ at this moment, it is also true that Joe* wants to choose, and does choose, *sell*. So Joe* wants and chooses to *buy* in period 1 and to *sell* in period 2. In allowing this sequence of actions, T_1 gives Joe* an opportunity to do something that he wants to do. If Joe* values opportunities to do as he wants, this feature of T_1 has value for him.

Consider how, at the end of period 2, Joe (the human being) might reflect on the actions he has taken. A conventional decision theorist might point out to him that he has acted on preferences that are not stable under experience and reflection, and that in consequence he has incurred an unambiguous loss – he has bought dear and sold cheap. Joe can concede this, yet still see both buying and selling as *his* autonomously chosen actions: he wanted to buy, and he bought; he wanted to sell, and he sold. He does not have to disown either of those actions as the work of an alien self, or as the result of weakness of will. While recognising that he has acted on unconsidered preferences, he can say that he has done what he wanted to do, when he wanted to do it. Without asserting that those preferences are his standard of value, he can say that he values the opportunity to act on them; and this can be a considered judgement on his part. All this becomes coherent if ‘he’ is understood as the continuing Joe*.

I now offer a sketch of how we might represent and generalise the idea that, for the continuing Joe*, T_1 is the most valuable of the three trees. The first step is to recognise that the standard of value for Joe* is not preference itself, but the opportunity to act on preferences, as and when they are felt. For my purposes, I do not need to use the information contained in the payoffs of Joe$_1$ and Joe$_2$. It is more useful to conduct the analysis in terms of *outcomes*. There are three relevant outcomes. One outcome, which results from *don’t buy*, is that Joe keeps his status quo level of wealth and doesn’t participate in the lottery. I denote this outcome *x*. The second possible outcome, denoted *y*, is that Joe doesn’t participate in the lottery and loses £2 of his status quo wealth; this results from *buy* followed by *sell*. The third possible outcome, denoted *z,* is that Joe gives up £9 of his status quo wealth and participates in the lottery; this results from *buy* followed by *hold*.

It is an important part of the problem that *y* is ‘unambiguously’ worse than *x*. One way of understanding this idea is to use the concept of *potential preference*.[12] To say that some preference is ‘potentially’ the preference of a particular agent is to say that the agent *might* have that preference. If an agent’s preferences are unstable, we can think of him at any given time as acting on some preference relation drawn (perhaps arbitrarily) from some fixed set of potential preference relations. If every potential preference relation ranks x above y, then x can be said to be unambiguously better than y.

To represent this idea, I do not need to model the set of potential preference relations explicitly, although I make the implicit assumption that each potential preference relation is complete and transitive (that is, that each of these relations is an ordering). Instead, I define a relation $\geq_*$ of *weak dominance* on the set of outcomes; $v \geq_* w$ is read as 'v weakly dominates w', and is interpreted as indicating that v is ranked at least as highly as w by every potential preference ordering. I stipulate that this relation is reflexive (that is, for every outcome v, $v \geq_* v$), but I do not require it to be complete. Given my interpretation of dominance, these properties are immediate implications of the assumption that each potential preference relation is complete and transitive. If v weakly dominates w but w does not weakly dominate v, I shall say that v *strictly dominates* w, denoted $v >_* w$. If each of v and w weakly dominates the other, I shall say that they are *dominance-equivalent*, denoted $v =_* w$.

I now need a notation for representing decision problems that are confronted by a single continuing agent but which extend over time. When dealing with decision problems which require single acts of choice at one moment in time, it is conventional to represent each such problem as a set of outcomes (the *opportunity set* or *menu*); the idea is that the agent is able to choose one outcome from this set. Of course, such a problem can also be represented as a decision tree with a single choice node, but representing it as a set of outcomes is simpler and more compact. It is possible to extend the set-theoretic notation to represent sequences of choices by using *nested sets*.[13] This idea is most easily explained by examples.

In this new notation, the decision problem that was previously represented by the tree T_1 is denoted by the nested set $S_1 = \{\{x\}, \{y, z\}\}$. The outer pair of curly brackets defines the choice facing the agent in period 1. This is a choice between the elements of the set that is specified by that pair of brackets – that is, the elements $\{x\}$, corresponding with *don't buy*, and $\{y, z\}$, corresponding with *buy*. Each of these elements is a (possibly degenerate) choice problem that will be confronted in period 2, if the relevant action is chosen in period 1. The singleton $\{x\}$ represents the fact that, if *don't buy* is chosen in period 1, there is no choice to be made in period 2, and the outcome will be x. The set $\{y, z\}$ represents the fact that, if *buy* is chosen in period 1, there will be a further choice to be made in period 2, between one action (*sell*) which leads to y and another (*hold*) which leads to z. Similarly, the decision problem previously represented by the tree T_2 is denoted by the nested set $S_2 = \{\{x\}, \{z\}\}$, while the decision problem previously represented by T_3 is denoted by the nested set $S_3 = \{\{x\}\}$.

Notice that, in this notation, each matched pair of curly brackets is associated with a specific time period; the succession of periods is represented by successively 'deeper' nesting of sets. Every outcome is nested within as many pairs of brackets as there are periods in the analysis (two in each of the variants of Joe's problem). The number of pairs of brackets within which each outcome is nested is the *depth* of the relevant nested set; thus S_1, S_2 and S_3 all have depth 2. The elements of each of these nested sets are themselves nested sets, but with depth 1. In general, a nested set of depth n (where $n > 1$) is a set of nested sets, each of which has depth $n – 1$; a nested set of depth 1 is a set of outcomes.

Using this notation, I now propose a principle for inducing a dominance relation among nested sets from the dominance relation among outcomes. I begin by considering nested sets of depth 1. I define a relation $\geq_*$ of weak dominance among such sets in the following way:

> *Dominance Extension Rule (for nested sets of depth 1)* For any nested sets R and S of depth 1: $R \geq_* S \Leftrightarrow (\forall v \in S)\ (\exists w \in R)\ w \geq_* v$.

Strict dominance and dominance equivalence are defined from weak dominance as before.

According to the Dominance Extension Rule, R weakly dominates S if every outcome in S is weakly dominated by some outcome in R. Thus, to say that R dominates S is to say that, for each potential preference ordering, for each outcome v in S, there is some outcome w in R such that w is at least as preferred as v. Given that dominance with respect to outcomes has been defined in terms of agreement among potential preferences, the Dominance Extension Rule is very natural.

For example, consider the outcomes x, y and z, as defined for Joe's problem. Recall that all potential preference orderings rank x strictly above y, but there is no agreement among these orderings about the ranking of x and z, or of y and z. Thus, x strictly dominates y, but there is no relation of dominance between x and z or between y and z. By the Dominance Extension Rule, $\{x, y\} =_* \{x\}$. Since y is weakly dominated by x, the best element of $\{x, y\}$ can be no better than the only element of $\{x\}$, whichever potential preference ordering we base our judgement on. In contrast, the same rule implies $\{y, z\} >_* \{z\}$. Since y is not weakly dominated by z, there is at least one potential preference ordering such that the best element of $\{y, z\}$ is better than z, while z obviously cannot be better than the best element of $\{y, z\}$.

It is straightforward to prove that, if the weak dominance relation $\geq_*$ is reflexive and transitive on the set of outcomes, it is also reflexive and transitive on the set of nested sets of depth 1, when that extension is as defined by the Dominance Extension Rule.[14] Because the formal properties of the weak dominance relation are preserved as we move from rankings of outcomes to rankings of nested sets of depth 1, we can use the same method to extend the weak dominance relation from nested sets of depth 1 to nested sets of depth 2, and so on indefinitely. Formally, I define:

> *Dominance Extension Rule (for nested sets of depth d > 1)* Let S and T be nested sets of depth d, where $d > 1$; notice that the elements of S and T are nested sets of depth $d - 1$. For all such S, T: $S \geq_* T \Leftrightarrow (\forall V \in T)\ (\exists W \in S)\ W \geq_* V$.

We can now apply the Dominance Extension Rule to the three nested sets of Joe's problem, that is $S_1 = \{\{x\}, \{y, z\}\}$, $S_2 = \{\{x\}, \{z\}\}$ and $S_3 = \{\{x\}\}$. First, compare S_2 and S_3. Both sets have the common element $\{x\}$, but S_2 has the additional element $\{z\}$. Since z neither dominates nor is dominated by x, the Dominance Extension Rule, applied at depth 1, implies that $\{z\}$ neither dominates nor is dominated by $\{x\}$. Thus, when the rule is applied at depth 2, we have $S_2 >_* S_3$. Now compare S_1 and S_2. Each of these sets contains two elements. One element, $\{x\}$, is common to both. The difference is that S_1 also contains $\{y, z\}$ while S_2 also contains $\{z\}$. I have already shown that the Dominance Extension rule, applied at depth 1, implies $\{y, z\} >_* \{z\}$. Thus, when the rule is applied at depth 2, we have $S_1 >_* S_2$. By transitivity, we have $S_1 >_* S_3$. On this analysis, the intuition of consumer sovereignty is vindicated: the absence of constraints on Joe's choices is unambiguously valuable to Joe.[15]

To see why the nested-set analysis delivers this result, consider the sequence of choices (*buy*, *sell*). Because this path leads to a dominated outcome (namely y), conventional analyses imply that the existence of this path has no positive value; if (because of Joe's inconsistency over time) it might in fact be chosen, its existence has negative value. The interests of the continuing Joe^M or Joe^S would then be best served by removing this apparently undesirable path, contracting the decision problem either to S_2 or to S_3. On the nested-set analysis, in contrast, this path is treated as the combination of two choices, each of which the continuing Joe* might want to make, and so has positive value. In period 1, Joe* has a choice between the actions *buy* and *don't buy*. *Don't buy* leads to $\{x\}$, while *buy* leads to $\{y, z\}$. Since neither of these sets dominates the other, each action is one that Joe* might want to take. The existence of each of these options, therefore, is valuable to Joe*. To remove the path leading to $\{y, z\}$, contracting the problem to S_3, would be to remove something of value. Now, suppose Joe* has chosen *buy* in period 1. In period 2, he faces a choice between y and z. Neither outcome dominates the other. Since each option is one that Joe* might wish to choose, the existence of each of them is valuable to him; to remove the path leading to y (contracting the problem to S_2) would again be to remove something of value.

6. Conclusion

Normative economics has been built on the assumption that each person has consistent and stable preferences, and has used these assumed preferences as the standard of value for that person. However, in the light of the recent findings of behavioural economics, it is no longer possible to treat this assumption either as a self-evident truth about human reason or as a well-established fact about how real economic agents think and act. This paper has asked whether it is possible to preserve the idea that there is value in respecting individuals' actual preferences, even if those preferences are inconsistent and unstable.

I have argued that each of us can value being free to act on his or her own preferences, considered or unconsidered, as and when we experience them. This is a kind of freedom that competitive markets are highly effective in providing, at least in relation to private goods and for individuals who are endowed with transferable goods that other people value.[16] Such a robust understanding of the value of opportunity may not be to everyone's taste, but I hope that at least I have persuaded the reader that those of us who do find it attractive can endorse it coherently and clear-sightedly.

Note on typescript

On p. 2, the combination of symbols [≥] is used in place of the standard symbol for 'weakly preferred to'.

Notes

1. I explain and defend this approach in Sugden (1989).

2. For an argument of this kind, see Broome (1991, pp. 10-12). Broome claims that it is a necessary truth that *all* relations of the form 'is at least as ... as' are transitive. In passing, I must report that I am not persuaded. Broome discusses the relation 'is at least as westerly as' which appears to be a counter-example. Broome tries to persuade us that appearances are deceptive, but in this case I think they are not.

3. Compare Broome's (1991, pp. 90-107) argument about the individuation of options and the need for 'rational requirements of indifference'.

4. See, for example, Starmer and Sugden (1989) and Hey and Orme (1994, p. 1296).

5. For surveys of the evidence, see Camerer (1995), Starmer (2000) and Kahneman and Tversky (2000).

6. Variants of the discovered preference hypothesis are proposed by Smith (1994) and Plott (1996); the term is Plott's. These hypotheses are discussed by Binmore (1999), Loewenstein (1999), Loomes (1999) and Starmer (1999).

7. Contrasting evidence on the effect of experience on anomalies can be found in Ariely et al (2003), List (2003) and Loomes et al (2003).

8. My hunch is that the analysis presented in Section 5 could be extended to include these cases by making use of the conventional multiple-self approach. Suppose that, at some point in a sequential decision problem, a person acts on a transient preference which, from the perspective of the 'continuing person', he disavows. That action can be treated as if it is made by another person; instead of a single-agent decision problem, we have a game with two players – the continuing person and an alien, transient self. There would be nothing especially original in such an analysis. The real challenge, as I see it, is to analyse cases in which the continuing person does *not* disavow his transient preferences.

9. Reviewing a range of investigations of the role of affect on decision-making, Isen (1999) concludes that positive affect tends to increase risk aversion in relation to decisions that are perceived to involve the possibility of serious loss.

10. It might be objected that the folding-back analysis attributes too much rationality to the agent. However, the outcome of the analysis is just the same if Joe is *myopic* – that is, if

each self acts as if its own preferences will determine behaviour at all subsequent nodes. The myopic Joe$_1$ will *buy*, expecting both Joe$_1$ and Joe$_2$ to *hold* (since that is the best action from Joe$_1$'s point of view); but while the myopic Joe$_1$ will indeed *hold*, the myopic Joe$_2$ will *sell*.

11. It has been suggested to me that the continuing Joe might have a preference for *spontaneity*, in the sense that he prefers to be the sort of person who acts on the preferences of the moment. This is not a metapreference in the standard theoretical sense of the term – that is, a higher-order preference *between alternative preference relations*. My hunch is that the most convincing analysis of a 'preference for spontaneity' would be very similar to my analysis of the value of consumer sovereignty.

12. I discuss the concept of potential preference in Sugden (1998).

13. The idea of representing decision problems as nested sets derives from Cubitt and Sugden (2001). However, the current analysis differs from that of Cubitt and Sugden by attributing decision nodes to specific time periods.

14. It is immediately obvious that if $\geq_*$ is reflexive on the set of outcomes, the Dominance Extension Rule implies that $\geq_*$ is reflexive on the set of nested sets of depth 1. To prove that transitivity is transmitted in a similar way, let Q, R and S be nested sets of depth 1, and suppose $Q \geq_* R$ and $R \geq_* S$. By the Dominance Extension Rule, $(\forall x \in S)\ (\exists y \in R)\ y \geq_* x$, and $(\forall y \in R)\ (\exists z \in Q)\ z \geq_* y$. Therefore $(\forall x \in S)\ (\exists z \in Q$ and $y \in R)\ (z \geq_* y$ and $y \geq_* x)$. Because $\geq_*$ is transitive on the set of outcomes, this implies $(\forall x \in S)\ (\exists z \in Q)\ (z \geq_* x)$. So by the Dominance Extension Rule, $Q \geq_* S$.

15. The reader may wonder if it would be even more valuable to Joe* to be able to choose which of the three decision problems to face. According to my analysis, the answer is 'No'. Since S_1 weakly dominates both S_2 and S_3, the Dominance Extension Rule applied at depth 3 implies $S_1 =_* \{S_1, S_2, S_3\}$. More generally: from the viewpoint of one's continuing self, an opportunity to impose constraints on oneself has zero value. This seems to be an inescapable implication of the conception of value that I am proposing. If constraints on a person's actions cannot have positive value for him, then an opportunity to impose constraints on oneself cannot have positive value either.

16. I argue this in Sugden (2004).

References

Ariely, Dan, George Loewenstein and Drazen Prelec (2003). ‘Coherent arbitrariness’: stable demand curves without stable preferences. *Quarterly Journal of Economics* 118: 73-105.

Binmore, Ken (1999). Why experiment in economics? *Economic Journal* 109: F16-F24.

Broome, John (1991). *Weighing Goods*. Oxford: Blackwell.

Camerer, Colin (1995). Individual decision making. In John Kagel and Alvin Roth (eds), *The Handbook of Experimental Economics*. Princeton, N.J.: Princeton University Press, pp. 587-703.

Cubitt, Robin and Robert Sugden (2001). On money pumps. *Games and Economic Behavior* 37: 121-160.

Gauthier, David (1986). *Morals by Agreement*. Oxford: Clarendon Press.

Hey, John and Chris Orme (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62: 1291-1326.

Isen, A.M. (1999). Positive affect. In T. Dalgleish and M. Power (eds), *Handbook of Cognition and Emotion*. London: Wiley and Sons.

Kahneman, Daniel and Amos Tversky, eds (2000). *Choices, Values and Frames*. Cambridge: Cambridge University Press.

List, John (2003). Does market experience eliminate market anomalies? *Quarterly Journal of Economics* 118: 41-71.

Loewenstein, George (1999). Experimental economics from the viewpoint of behavioural economics. *Economic Journal* 109: F25-34.

Loomes, Graham (1999). Some lessons from past experiments and some challenges for the future. *Economic Journal* 109: F35-F45.

Loomes, Graham, Chris Starmer and Robert Sugden (2003). Do anomalies disappear in repeated markets? *Economic Journal* 113: C 153-166.

Plott, Charles (1996). Rational individual behaviour in markets and social choice processes: the discovered preference hypothesis. In Kenneth J. Arrow, Enrico Colombatto, Mark Perlman and Christian Schmidt (eds), *The Rational Foundations of Economic Behaviour*. Basingstoke: Macmillan, pp. 225-250.

Smith, Vernon (1994). Economics in the laboratory. *Journal of Economic Perspectives* 8: 113-131.

Starmer, Chris (1999). Experimental economics: hard science or wasteful tinkering? *Economic Journal* 109: F5-F15.

Starmer, Chris (2000). Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38: 332-382.

Starmer, Chris and Robert Sugden (1989). Violations of the independence axiom in common ratio problems: an experimental test of some competing hypotheses. *Annals of Operations Research* 19: 79-102.

Sugden, Robert (1989). Maximizing social welfare: is it the government's business? In Alan Hamlin and Philip Pettit (eds), *The Good Polity* (Oxford: Basil Blackwell).

Sugden, Robert (1998). The metric of opportunity. *Economics and Philosophy* 14: 307-337.

Sugden, Robert (2004). The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. *American Economic Review* 94: 1014-1033

Figure 1: Joe's problem, unconstrained (Tree T_1)

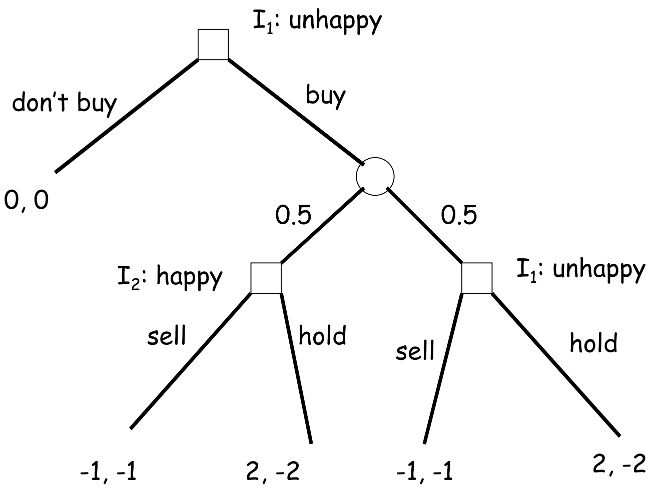


Figure 2: Joe's problem with a constraint (Tree T_2)

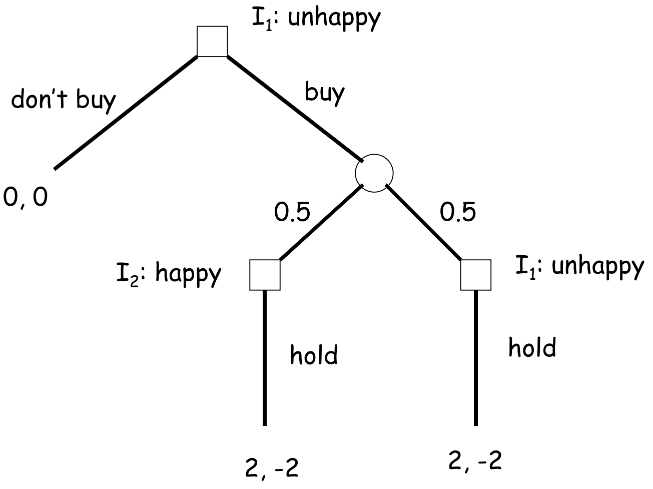
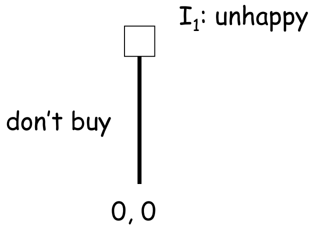


Figure 3: Joe's problem with a different constraint (Tree *T_3*)


I1: unhappy
don't buy
0, 0