# Transitivity, the Sorites Paradox, and Similarity-Based Decision-Making

Version of January 22, 2005.


Alex Voorhoeve
Department of Philosophy,
Logic, and Scientific Method
London School of Economics
Houghton Street
London WC2A 2AE, UK
a.e.voorhoeve@lse.ac.uk

and

Ken Binmore
Economics Department
University College London
Gower Street
London WC1E 6BT, UK
k.binmore@ucl.ac.uk

Correspondence should be directed to Alex Voorhoeve


Word count (including footnotes and references): Approx. 5,000

# Transitivity, the Sorites Paradox, and Similarity-Based Decision-Making

**Abstract:** A persistent argument against the transitivity assumption of rational choice theory postulates a repeatable action that generates a significant benefit at the expense of a negligible cost. No matter how many times the action has been taken, it therefore seems reasonable for a decision-maker to take the action one more time. However, matters are so fixed that the costs of taking the action some large number of times outweigh the benefits. In taking the action some large number of times on the grounds that the benefits outweigh the costs every time, the decision-maker therefore reveals intransitive preferences, since once she has taken it this large number of times, she would prefer to return to the situation in which she had never taken the action at all.

We defend transitivity against two versions of this argument: one in which it is assumed that taking the action one more time never has any perceptible cost, and one in which it is assumed that the cost of taking the action, though (sometimes) perceptible, is so small as to be outweighed at every step by the significant benefit. We argue that the description of the choice situation in the first version involves a contradiction. We also argue that the reasoning used in the second version is a form of similarity-based decision-making. We argue that when the consequences of using similarity-based decision-making are brought to light, rational decision-makers revise their preferences. We also discuss one method that might be used in performing this revision.

## 1 Introduction

A persistent argument against the transitivity assumption of rational choice theory postulates a repeatable action that generates a significant benefit at the expense of a negligible cost. No matter how many times the action has been taken, it therefore seems reasonable for a decision-maker to take the action one more time. However, matters are so fixed that the costs of taking the action

some large number of times outweigh the benefits. In taking the action some large number of times on the grounds that the benefits outweigh the costs every time, the decision-maker therefore reveals intransitive preferences, since once she has taken it this large number of times, she would prefer to return to the situation in which she had never taken the action at all.

A defender of transitivity must insist that the decision-maker has made a mistake somewhere, and should therefore revise at least one of her pairwise preferences. This defender should ideally be able to help with this revision by locating the mistake in either the conception of the decision situation or in the process of evaluation that leads to the irrational pairwise preferences. If the mistake lies in the process of evaluation, then the defender should ideally be able to offer both an explanation of the mistake, and an alternative procedure by which it can be avoided.

In this paper, we try to do just this. In section 2, we discuss a version of the argument in which it is assumed that taking the action one more time never has any perceptible costs. Taking Warren Quinn's (1990) 'Paradox of the Self-Torturer' as an example, we argue that the description of the choice situation supposed in this version of the argument is mistaken.

In section 3, we turn to a version of the argument that assumes that the cost of taking the action, though (sometimes) perceptible, is so small as to be outweighed at every step by the significant benefit. This version of the argument is advanced by Erik Carlson (1996) in a revision of Quinn's paradox. We show that the reasoning used in this version of the argument is a form of similarity-based decision-making (see Tversky, 1969, 1977; Rubinstein, 1988, 2003). Real people certainly reason in this way. However, we argue that the preferences revealed by this form of decision-making are not fully considered. That is to say, when the consequences of using similarity-based decision-making in the cases under discussion are brought to light, rational decision-makers would revise their preferences. We also discuss one of many methods that might be used in performing this revision.

## 2 Imperceptible Differences and the Sorites Paradox

In Quinn's 'Paradox of the Self-Torturer', Alice is strapped to a conveniently portable machine that administers a continuous electric current, the strength of which depends on the position of a dial with 1,001 notches, numbered 0 to 1,000. Alice is first allowed as much time as she likes to experiment freely with the dial, administering shocks of various intensity to herself to find out how she reacts to different settings of the dial. Quinn postulates that the results of her experimentation are as follows: she can tell the difference between notches that are far enough apart, but finds that she "cannot feel any difference in comfort between adjacent settings", because "we have made the increments of current too small to make *any* difference in comfort." (1990, p. 79 and p. 81.) It is further assumed that the current has no effects on Alice's well-being other than the discomfort it causes.

After experimenting, Alice is offered the following deal. Starting with the dial at its lowest notch, she may advance the dial one notch once a week. Each time she advances the dial by one notch, she is paid $10,000. But she can never reverse the process and return to an earlier notch. If she eventually finds the pain hard to bear, she must nevertheless continue to endure the discomfort for the rest of her life.

Quinn argues that at every notch, Alice appears to have a good reason to turn the dial, because she feels no worse by turning the dial, but reaps a substantial financial reward. But, so Quinn supposes, the result of turning the dial to its highest setting is that Alice must suffer a pain so excruciating she would be willing to return all the money she has earned to return the dial to its lowest setting. She thereby reveals that her considered preferences are intransitive. However, Quinn remarks, we cannot simply dismiss Alice's preferences as irrational because they are intransitive (1990, p. 80). For, so Quinn argues, given the description of the choice situation, it seems entirely natural and appropriate for her to have such preferences. Nonetheless, if she proceeds on the basis of her pairwise preferences between notches, as Quinn believes the standard theory of rational choice tells her she must, she ends up in a bad situation.

Quinn is wrong to suppose that if Alice had intransitive preferences in this case, a rational choice

theorist would tell her to advance the dial at each notch. Instead, since the standard theory considers intransitive preferences irrational, he would advise Alice to reconsider her preferences. Furthermore, we believe Quinn is mistaken to suppose that Alice will have intransitive preferences for the reason he offers.

To see why, imagine that during her period of experimentation, Alice asks herself how she feels at each notch and records her responses, using terms like 'no discomfort', 'just slightly uncomfortable, maybe', 'mild pain', 'great pain', etc. (see also Arntzenius and McCarthy, 1997). We will refer to these descriptions of discomfort as 'levels of discomfort'. If she repeatedly returns to the same notch, or just records her feelings at regular intervals while at the same notch, it is possible that on different occasions she will feel differently at the same notch, due to unknown or random processes in her nervous system. In this case, she will represent the information about 'how a notch feels' by a distribution over different levels of discomfort. We will say that two notches 'feel the same' or 'are indistinguishable' to Alice when these recorded distributions are identical.

It is clear that on this understanding of 'feels the same as' or 'is indistinguishable from', it cannot be true that all adjacent notches feel the same when she runs through all notches in ascending order. For the assumption that Alice feels no discomfort at notch 0, together with the assumption that all adjacent notches feel the same, yields the conclusion that Alice feels no discomfort at notch 1, and so on for all notches until notch 1000. This contradicts the assumption that she is in excruciating pain at that notch. If Alice wishes to maintain the assumptions about the pain at notch 0 and notch 1000, she therefore cannot assume that as she runs through all notches in ascending order, advancing the dial by one notch will never make any difference in comfort.

This reasoning is independent of how small we make the increments in current. This means that for a sequence of this sort, we must reject the popular idea that there is a "least-noticeable difference": a magnitude of physical change so small that human beings always fail to detect a difference between situations in which a change smaller than this magnitude has and hasn't occurred. This idea was first concisely incorporated into a model of decision-making by Duncan Luce (1956), who argued

that "the imperfect powers of discrimination of the human mind" meant that inequalities in physical states become recognizable only when they are of sufficient magnitude. However, contrary to what Luce seems to suggest, the fact that we must reject the existence of least-noticeable differences for a sequence of changes of the kind Alice faces when she runs through all notches in ascending order does not imply that we are attributing to Alice unlimited capacities of discrimination. For consider a device that registers charge only in whole kilovolts. If we hooked this device to a machine that administered a current of varying voltage, started with the dial at 0 and kept increasing the charge by small increments, then at some point the device will change from registering '0 kilovolt' to registering '1 kilovolt'. This implies that there are no just-noticeable differences in the sense under discussion for this device, even though its capacities of discrimination are limited.

Nonetheless, we can accommodate some kinds of indistinguishability that make use of the idea of just-noticeable differences in our description of Alice's situation, and perhaps this is what Luce had in mind. One is that Alice might find any two adjacent notches indistinguishable in isolation. By this, we mean that Alice only compares 'how it feels at notch $n$ after having experienced notch $n+1$' with 'how it feels at notch $n+1$ after having experienced notch $n$.' Two adjacent notches might be indistinguishable in this way because Alice's pain experience at a particular notch might (because of some neurophysiological process that we need not understand) depend on the current she was exposed to before. Thus, it might be the case that if the previous current is very different, she experiences the current at notch $n$ in one way, but if it is similar (i.e. the difference between them is smaller than the just-noticeable difference), she experiences it in another way. To illustrate, suppose that if it directly follows notch 0, Alice reports at notch 750 that she feels 'great pain' 51 percent of the time, and 'very great pain' 49 percent of the time. At notch 751, if it directly follows notch 0, Alice reports that she feels 'great pain' 49 percent of the time, and 'very great pain' 51 percent of the time. But if notch 750 follows notch 751, or notch 751 follows 750, she reports 'great pain' and 'very great pain' exactly half the time. If these assumptions hold, and if Alice confines her comparison of the two notches to the information generated by a sequence of '751 after 750' and

'750 after 751', then these two notches will be indistinguishable to her in isolation.

We can also permit what one might call 'introspective indistinguishability'. Suppose Alice does not tabulate her experiences as we supposed, but instead relies only on her memory for the comparison of her experience at different notches. It is conceivable that when Alice is presented with two similar stimuli in succession, in her memory the first stimulus always becomes assimilated to the second, so that she finds them introspectively indistinguishable.

These forms of indistinguishability rely on contextual dependence of perception or inconstancy in our memory. In this, we follow some recent discussions of phenomenal sorites (see Graff 2001, Mills 2001, Regan 2000, and Raffman 1994). Some of the attraction of the idea of indistinguishability may simply lie in the fact that if the number of notches is far larger than the number of levels of discomfort that can be experienced over the same range of current, then Alice may well find that neighbouring notches often feel the same. However, we suggest that its attractiveness may also have something to do with a tendency to focus on isolated pairwise comparison, or introspective comparison when imagining Alice's case.

Both forms of indistinguishability are compatible with the fact that as she runs through the entire sequence of notches in ascending order from 0 to 1000, the probability of Alice experiencing a certain level of discomfort changes. For that notch 750 and 751 are indistinguishable in isolation does not mean that the probability that Alice will experience any given kind of discomfort remains unchanged when she moves from the distribution associated with 'notch 750 after notch 749' to the distribution associated with 'notch 751 after notch 750' by turning the dial from notch 750 (after notch 749) to notch 751. Nor would the fact that Alice's memory always deceives her into thinking she felt the same at notch 750 as she feels now at notch 751 entail that she always did feel the same.

With this in mind, we suggest Alice will interpret the data generated by her limited period of experimentation as follows. Alice will see each notch as characterized by a probability distribution over different levels of discomfort and by a monetary prize (see Arntzenius and McCarthy, 1997). This probability distribution will be her estimate of the probability that at any given time she will

| | **Notch 10** | **Notch 11** |
|---|---|---|
| Money | \$100,000 | \$110,000 |
| Expected pain | | |
| Probability of level of discomfort 0 | 0.82 | 0.81 |
| Probability of level of discomfort 1 | 0.18 | 0.19 |

Table 1: A hypothetical representation of notches 10 and 11

experience a particular level of discomfort while at that notch, given the previous notch she was exposed to.

For example, suppose at notch 10 she will have \$100,000 and that the probabilities she attaches to being at level of discomfort 0 ('no discomfort') and 1 ('slight discomfort') at any given time are 0.82, and 0.18, respectively. At notch 11 she will have \$110,000, while the respective probabilities are 0.81 and 0.19 (see Table 1). (These probabilities can be made conditional on the previous notch Alice was exposed to. For simplicity, we leave this aside in what follows, but the probabilities we discuss for any notch $n$ can be seen as conditional on experiencing that notch after being at notch $n-1$, so that they represent the probabilities that Alice faces when she goes through the notches in ascending order from 0 to 1000.) She will then determine her preference between notch 10 and notch 11 by asking herself whether, starting from the situation at notch 10, it is worth increasing the likelihood of being in state 1 at any point in time by 1 percent to get an extra \$10,000.

In sum, since the probability distribution over levels of pain must differ from its predecessor for at least one notch as Alice runs through notch 0 to 1000 in sequence, the 'imperceptible difference' argument collapses.

# 3 Similarity-Based Decision-Making and a Revised Version of Quinn's Paradox

There is, however, a revised version of Quinn's paradox in which Alice is supposed to reason as follows: "Even if some adjacent settings are different in terms of expected discomfort, the increase in expected discomfort is at most only very slight. But having an extra $10,000 is a great benefit. So I should increase the setting by one." (See Carlson, 1996; Arntzenius and McCarthy, 1997, p. 138.)[1] Nonetheless, the combined effect of all the increases in discomfort is assumed to outweigh the total amount of money to be gained.

The form of decision-making that Alice is applying in this case is what Amos Tversky (1969, 1977) and Ariel Rubinstein (1990, 2003) have called "similarity-based decision-making". Tversky and Rubinstein hypothesize that people use this form of decision-making to simplify their choice between multidimensional alternatives. Rubinstein's characterization of this form of decision-making is as follows. When deciding between multi-dimensional alternatives, say bundles of (expected) pain and money $(p_i, m_i)$ and $(p_j, m_j)$, a decision-maker goes through the following three-stage procedure.

**Stage 1:** The decision-maker looks for dominance. If $p_i < p_j$ and $m_i > m_j$, then bundle $(p_i, m_i)$ is preferred over bundle $(p_j, m_j)$.

**Stage 2:** The decision-maker looks for similarities between $p_i$ and $p_j$ and between $m_i$ and $m_j$. If she finds similarity in one dimension only, she disregards this dimension, and determines her preference between the two pairs using only the dimension in which there is no similarity. For example, if $p_i$ is similar to $p_j$ but $m_i$ is not similar to $m_j$, and $m_j > m_i$, then bundle $(p_j, m_j)$ is preferred over bundle $(p_i, m_i)$.

**Stage 3:** If the first two stages were not decisive, the choice is made using a different criterion.

[1] Arntzenius and McCarthy believe Alice is mistaken to reason in this way, as do we. However, they do not offer our diagnosis of the kind of reasoning employed, nor do they discuss our proposed way in which Alice can revise her preferences.

Tversky and Rubinstein find evidence supporting the hypothesis that people make decisions in this way in a variety of experiments involving gambles (with probability and prizes as dimensions of the alternatives), the choice of applicants (where the dimensions were taken to be 'intellectual ability', 'emotional stability' and 'social facility'), and intertemporal tradeoffs (where the dimensions were time and money). They also demonstrate that this decision-making procedure can easily generate intransitive preference orderings.[2] Indeed, Tversky systematically managed to get subjects who used similarity-based decision-making to reveal intransitive preferences by presenting them with a sequence of pairwise choices between multi-dimensional alternatives, in which each member of the sequence differed very little from its predecessor along the first dimension, but differed significantly from its predecessor along at least one other dimension. The first dimension was chosen so as to be generally "more important" than the other dimensions, at least when the difference between alternatives along this dimension was significant. In choices between neighbouring pairs in the sequence, subjects made the non-similar dimension or dimensions decisive. However, in choices between the first and last members of the sequence, there were no similarities along any dimensions, and the difference between these alternatives along the first dimension became the decisive factor. In cases where subsequent members of the sequence got slightly better along the first dimension, but significantly worse along the other dimension or dimensions, this led to earlier members of the sequence being preferred to their neighbours in the sequence in pairwise comparison, but the last member of the sequence being preferred to the first, generating an intransitive ordering.

We can illustrate this pattern of choice by the following example, adapted from Tversky (1967, p. 32.) Consider a situation in which three alternatives, $x$, $y$, and $z$, vary along two dimensions, I and II, as depicted in Table 2.

[2] Rubinstein (1990) demonstrates that the use of similarity relations in decision making over two-dimensional alternatives may be consistent with a transitive ordering, but only under very restrictive conditions on the type of similarity relation used and on the criterion used in Step 3 of the procedure. Xavier Vilà (1998) strengthens this result by showing that any attempt to order alternatives of three or more dimensions using similarity relations will generate intransitivities.

| | | *Dimensions* | |
|---|---|---|---|
| | | I | II |
| | $x$ | $2\epsilon$ | $6\epsilon$ |
| *Alternatives* | $y$ | $3\epsilon$ | $4\epsilon$ |
| | $z$ | $4\epsilon$ | $2\epsilon$ |

Table 2: A payoff matrix that can generate intransitive preferences

The alternatives might be job applicants varying in intelligence (I) and experience (II), where the entries are the candidates' scores on the corresponding dimensions. Suppose the subject uses the similarity-based procedure defined above to choose between each pair of alternatives, where two alternatives are judged to be similar along a dimension if the difference between the alternatives' scores on that dimension is less or equal to $\epsilon$. Suppose further that stage 3 of this procedure is as follows: if the first two stages were not decisive, choose the alternative with the highest score on Dimension I. Since $x$ and $y$ and $y$ and $z$ are similar along the first dimension, the choice is made on the second dimension, and the subject chooses $x$ over $y$ and $y$ over $z$. But since there is no similarity between $x$ and $z$, the subject moves to stage 3 to decide between them and chooses $z$ over $x$, revealing an intransitive chain of preferences.

It is easy to see how Alice's reasoning in the revised version of Quinn's paradox fits this procedure. In comparing two adjacent notches, Alice is hypothesized to always regard the expected pain as 'similar', and the money as 'not similar'. (For example, in the hypothetical case described in Table 1, Alice would regard the pain represented by the probabilities (82; 18) and (81; 19) over pain levels (0; 1) as similar, and the amounts \$100,000 and \$110,000 as dissimilar.)[3] In accordance with

[3] As the amount of money Alice has won increases, \$10,000 may cease to be a significant difference. That is, when Alice is comparing notches 990 and 991, she is comparing prizes of \$9,900,000 and \$9,910,000. These may also seem similar, if Alice (sometimes) employs a ratio-similarity relation, rather than an absolute difference similarity relation. (For discussion of these types of similarity relations, see Rubinstein, 1990.) This could be remedied by changing the example so that Alice gets more money for turning the dial at higher notches.

stage 2 of the procedure, she opts for the latter of the two notches. However, the expected pain, which is systematically disregarded in pairwise comparisons between adjacent notches, emerges as a significant and decisive factor when the first and last notches are compared with each other.

What are we to make of the intransitivities so generated? Is this behaviour necessarily irrational? Similarity-based decision-making yields an orderly and recognizable way of decision-making. It also simplifies decision-making in several ways. Firstly, it focuses on differences rather than absolute values. This makes it compatible with what has been called a basic principle of perception and judgement: we are better attuned to the evaluation of changes or differences than to the evaluation of absolute magnitudes (Kahneman and Tversky, 1979). Second, if one alternative is slightly better than another along all relevant dimensions, it will be immediately apparent in intradimensional comparison and the choice will be easy. By contrast, if the alternatives are first evaluated independently as a whole, this dominance relation might be obscured, and the decision would in any case be made only after the possibly complex 'total evaluation' of the alternatives (Tversky, 1967, p. 42). Finally, by placing intradimensional evaluation (in stage 1 and stage 2) before the possible use of interdimensional evaluation (in stage 3), the procedure makes use of the fact that intradimensional evaluation is simpler, because the compared quantities are expressed in the same units (Tversky, 1967, p. 43). For example, when Alice applies the procedure to the comparison of adjacent notches, if she judges the expected pain at neighbouring notches to be similar, and the money at neighbouring notches to be dissimilar, she never really has to ask herself whether $10,000 is worth the possible extra discomfort. She simply compares pain with pain, and money with money. The procedure's prevalence can therefore be explained by its saving on decision-making costs and by the fact that people do not know that it may lead to violations of transitivity, because most environments in which it is applied are not designed to take advantage of its weaknesses. Similarity-based decision-making cannot, therefore, be dismissed out of hand as irrational in all decision-making environments.

Nonetheless, in environments that are designed to take advantage of the procedure's tendency to generate intransitive preferences, like Alice's imagined situation, it is a mistake to employ it so

long as an alternative method which yields reliable, transitive orderings is available, and the costs of employing this method are not prohibitive. Alice has good reason to regard her initial ordering as mistaken, since in addition to the familiar problems that intransitive orderings generate, she will end up badly by adhering to it. And she has another reason to regard her initial ordering with suspicion: it will not be stable across different descriptions of the same decision problem. (This means she will violate what Kahneman and Tversky (1984) call "the principle of invariance".) For imagine that the decision problem is presented to her as follows. She is asked to consider for each notch $n$ how much she would need to receive (or be willing to pay) while at notch 0 to render her indifferent between being at notch $n$ (with its concomitant probability distribution over levels of pain and its sum of money) and being at notch 0. To prevent the use of similarity-based reasoning, she is asked to perform this evaluation in an order that ensures that each notch she is comparing to her 'baseline' of notch 0 is sufficiently different from the previous notch she compared to notch 0. A notch $n$ is then better than another notch $k$ if Alice would need to receive more (or pay less) while at notch 0 to render her indifferent between staying at notch 0 and switching to notch $n$ than she would need to render her indifferent between staying at notch 0 and switching to notch $k$. For example, suppose that Alice would have to receive \$60,000 to make her indifferent between staying at notch 0 and moving to notch 11, at which she receives \$110,000 and faces a probability of experiencing no discomfort of 0.81 and of 'slight discomfort' of 0.19. Suppose further that she would be willing to pay at most \$50,000 to stay at notch 0 rather than move to notch 749, at which she receives \$7,490,000 and has a probability of being in 'great pain' of 0.51 and of being in 'very great pain' of 0.49. Then notch 11 is better than notch 749. The best notch is the notch for which she would need to receive the highest amount to render her indifferent between staying at notch 0 and switching to that notch.[4] Since this method assigns a numerical value to each notch and hence necessarily

[4] This procedure will not work for notches close to notch 0. Should she have any reason to think any of these notches might be the best notch, then Alice should compare the notches close to notch 0 and any other candidates for 'best notch' with a notch sufficiently far removed from all of them. She should then determine how much she would need to receive to render her indifferent between staying at that notch and moving to each of the candidate notches.

respects transitivity, the ordering that results will be different from Alice's initial ordering.

It will therefore be clear to Alice that her initial pairwise preferences are mistaken. People generally share this view: when Tversky's experimental subjects were told that they had revealed intransitive preferences, they typically saw themselves as "having made a mistake somewhere" (Tversky, 1967, p. 40). It may not be clear to them, however, where they made the mistake or how to correct it. When presented with the same sequence of pairwise choices a second time, in full knowledge of the fact that their initial choices generated an intransitive ordering, they might still feel drawn to choose in the same way. What is someone to do who accepts the normative force of the principle of transitivity and the principle of invariance but who is also drawn to using similarity-based decision-making?

We suggest that such a person should regard her pairwise preferences from the initial choice situation as an artifact of the arrangement of the alternatives, which led her to systematically undervalue an important aspect of the alternatives. Instead, she should determine her preferences by a reliable method that bypasses stage 2 of the procedure. One possible method of this kind would be the one just outlined. Alice could test the reliability of the ordering elicited by this method by checking whether it agrees with her decisions in other presentations of the same decision problem (excluding, of course, presentations that would tempt her to employ similarity-based reasoning). For example, she could check whether it agrees with her pairwise preferences in a direct choice between notches that are far enough apart. Thus, to confirm the preference for notch 11 over notch 749 that she found by the indirect method of asking how much she would need to receive (or pay) while at notch 0 to render her indifferent between each of these notches and notch 0, she could ask herself whether she would prefer notch 11 to notch 749 in a direct comparison.[5] She could also check whether she would arrive at the same preference ordering no matter which notch she used as

[5] Empirical evidence indicates that the preferences expressed between pairs that are far enough apart will be transitive. (See Tversky, 1967, pp. 36-37.) Though this doesn't imply that the direct pairwise comparison of notches that are far enough apart will yield the same ordering as our suggested method, it ensures that it is at least not a foregone conclusion that the two will conflict.

a baseline.[6] To illustrate, Alice's preference for notch 11 over notch 749 with notch 0 as the baseline would be consistent with the preferences she revealed using notch 1000 as a baseline if she would be indifferent between, say, getting another $2,000,000 on top of her $10,000,000 while at notch 1000 and moving to notch 11, while she would also be indifferent between, say, getting another $500,000 on top of her $10,000,000 while at notch 1000 and moving to notch 749. Notch 11 would then be better than notch 749 when judged from a baseline of notch 1000, just as it was when judged from a baseline of notch 0.

It may well be that her choices in some of these different presentations of the same decision problem will differ. Should there be such inconsistencies, then Alice should return to those judgement that differed between different presentations and ask herself which of these judgments she trusts more. This may involve inquiring into the process of evaluation she is employing under each presentation, and asking whether it is a process that she thinks is dependable. Sometimes she will come to regard the ordering resulting from a particular presentation as undependable, as we have proposed she should do with the ordering elicited by the initial presentation of the decision situation. By a process of jockeying–making judgments under different presentations, checking their consistency, questioning inconsistent judgments and their grounds, discarding orderings resulting from undependable presentations and methods of evaluation in some cases, revising her judgments in others, again checking their consistency, etc.–she should ultimately arrive at an ordering that is consistent across different (non-misleading) presentations and that respects transitivity.[7]

[6] Keeping in mind the caveat about notches close to the baseline mentioned in note 4.

[7] An example of this process is Leonard Savage's (1972, pp. 101-04) response to a case proposed by Maurice Allais (1953). Allais elicited Savage's on the spot preferences in two decision situations each involving two gambles, and showed that his choices in these situations, taken together, contradicted the postulates of Savage's theory. Though he confessed that he was still intuitively attracted to his initial preferences, Savage revised his preferences by considering a different presentation of the same decision situations which he regarded as superior to the first. He saw this procedure as correcting an error in his initial preferences. See also the description of the process of "jockeying" to arrive at consistent prior subjective estimates of probability distributions in Luce and Raiffa (1957, pp. 299-302).

## Acknowledgements

We thank Luc Bovens, John Broome, Erik Carlson, Marco Mariotti, Michael Otsuka, Wlodek Rabinowicz, Stuart Rachels and Ariel Rubinstein for helpful comments. This paper was presented at the London University Graduate Conference in January 2004 and to the LSE Choice Group in March 2004. We thank those present at these meetings for their comments. Alex Voorhoeve's work on this article was supported by the Analysis Trust.

## References

Allais, M.: 1953, 'Le comportement de l'homme rationnel devant le risque: Critique des postulats et axioms de l'école Americaine,' *Econometrica* 21: 503-46.

Arntzenius, F. and D. McCarthy: 1997, 'Self Torture and Group Beneficence,' *Erkenntnis* 47: 129-44.

Carlson, E.: 1996, 'Cyclical Preferences and Rational Choice,' *Theoria*, 62: 144-60.

Graff, D.: 2001, 'Phenomenal Continua and the Sorites,' *Mind* 110: 905-35.

Kahneman, D. and A. Tversky: 1979, 'Prospect Theory,' *Econometrica* 47: 263-91.

Kahneman, D. and A. Tversky: 1984, 'Choices, Values, and Frames,' *American Psychologist* 39: 341-50.

Luce, D.: 1956, 'Semiorders and a Theory of Utility Discrimination,' *Econometrica* 24: 178-91.

Luce, D. and H. Raiffa: 1957, *Games and Decisions*, John Wiley, New York.

Mills, E.: 2001, 'Fallibility and the Phenomenal Sorites,' *Noûs* 36: 384-407.

Quinn, W.: 1990, 'The Paradox of the Self-Torturer,' *Philosophical Studies* 59: 79-90.

Raffman, D.: 1994, 'Vagueness Without Paradox,' *Philosophical Review* 103: 41-74.

Regan, D.: 2000, 'Perceiving Imperceptible Harms: With Other Thoughts on Transitivity, Cumulative Effects, and Consequentialism,' in M. Almeida, (ed.), *Imperceptible Harms and Benefits*, Kluwer Academic Publishers, Dordrecht, pp. 49-73.

Rubinstein, A.: 1990, 'Similarity and Decision-Making Under Risk (Is There a Utility Theory Resolution of the Allais Paradox?),' *Journal of Economic Theory* 46: 145-53.

Rubinstein, A.: 2003, 'Economics and Psychology? The Case of Hyperbolic Discounting,' *International Economic Review* 44: 1207-16.

Savage, L.: 1972, *The Foundations of Statistics, second, revised edition*, Dover, New York.

Tversky, A.: 1967, 'Intransitivity of Preferences,' *Psychological Review* 84: 31-48.

Tversky, A.: 1977, 'Features of Similarity,' *Psychological Review* 84: 327-52.

Vilà, X.: 1998, 'On the Intransitivity of Preferences Consistent with Similarity Relations,' *Journal of Economic Theory* 79: 281-87.