Behavioural Public Policy (2025), **9**, 285–292
doi:10.1017/bpp.2024.39

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Daniel Kahneman and the concept of the true self

Robert Sugden

University of East Anglia, UK
Email: r.sugden@uea.ac.uk

(Received 31 July 2024; accepted 2 August 2024)

Abstract

Kahneman and Tversky's 'Prospect Theory' paper famously demolishes expected utility theory as a predictive device. However, it presents deviations from that theory as 'normatively unacceptable' and argues that decision-makers would normally correct them when possible. In a later paper, Kahneman rejects a similar argument (the 'discovered preference' hypothesis) advanced by Plott. Later still, Kahneman endorses Sunstein and Thaler's 'libertarian paternalism', which aims to help people avoid deviating from their 'true' preferences. I report an email correspondence between Kahneman and me in which we debated whether his position on libertarian paternalism was consistent with his critique of Plott's hypothesis.

Indisputably, Daniel Kahneman and Amos Tversky's (1979) paper 'Prospect Theory' is a – perhaps *the* – founding text of behavioural economics as a sub-discipline. It was a systematic compilation of existing knowledge about patterns of human decision-making that contravene the axioms of expected utility theory (EUT), backed up by a battery of evidence from the authors' own experiments. It proposed a unified theory of choice under risk that could explain those patterns. Its publication in *Econometrica* (a peak of achievement for any economist) legitimated as economic science both the use of experiments to test decision theories and the development of non-EUT theories to explain anomalous experimental observations.

The 1970s and 1980s were probably the high water mark of rational choice theory as a tool of economics, political science and philosophy. Tempting though it is to say that Kahneman and Tversky's paper was a bombshell, a better military metaphor would be a D-Day landing. Much more than other psychologists of the time, Kahneman and Tversky were not satisfied to criticise economics from outside. They wanted their ideas to be recognised *as economics*, and their attack was carefully planned to achieve this objective. It succeeded.

My concern in the current paper is with a short paragraph in 'Prospect Theory' (the *concession paragraph*) that seems out of keeping with the rest of the paper. Having presented a devastating array of evidence of violations of EUT, and having

proposed a psychologically based theory to explain this, Kahneman and Tversky make a surprising concession to rational choice theory. Noting that prospect theory assumes that values are attached to changes rather than to final states and that decision weights do not coincide with stated probabilities, they say:

> These departures from expected utility theory must lead to normatively unacceptable consequences, such as inconsistencies, intransitivities, and violations of dominance. Such anomalies of preference are normally corrected by the decision maker when he realizes that his preferences are inconsistent, intransitive, or inadmissible. In many situations, however, the decision maker does not have the opportunity to discover that his preferences could violate decision rules that he wishes to obey. In these circumstances the anomalies implied by prospect theory are expected to occur. (p. 277)

The first sentence asserts the normativity of the axioms of decision theory that prospect theory does not satisfy. It is followed by two strong empirical hypotheses – that real decision-makers *want* to obey those axioms and that when they discover preferences that contravene them, they *correct* those preferences. Kahneman and Tversky are hypothesising a form of *latent rationality*. They are proposing that individuals whose decisions contravene EUT are acting contrary to their own 'true' preferences – preferences that already exist in some undiscovered form, or that the individual is committed to constructing but has yet to construct. Puzzlingly, given that 'Prospect Theory' is a fundamentally empirical paper, no justification is given for the normative claim in the concession paragraph, and no evidence is presented in support of its two empirical hypotheses. All three are stated as if they were self-evident truths.

In understanding Kahneman's work as a whole, how much significance should we attach to this concession? Did Kahneman effectively withdraw it when behavioural economics had become a recognised sub-discipline but was still being criticised by rational choice theorists? When, later still, he endorsed Cass Sunstein and Richard Thaler's approach to normative behavioural economics, was he implicitly invoking an assumption of latent rationality? These are the topics of my paper. But, rather than proposing and defending answers to these questions, I will allow Kahneman to speak for himself – in a little-known comment on a paper by Charles Plott, and in an email correspondence with me.

Kahneman and the discovered preference hypothesis

On reading the concession paragraph when 'Prospect Theory' was first published, my immediate thought was that it was probably a reluctant response to a critical referee's insistence on the normativity of EUT. Later it occurred to me that it might have been a deliberate tactical move by Kahneman and Tversky: anticipating the criticism that their theory was normatively unacceptable and they might have chosen to restrict their initial attack on decision theory to a single front.

In any event, as experimental evidence of violations of EUT accumulated in the 1980s and 1990s, many economists defended EUT by using essentially the same argument as in Kahneman and Tversky's concession, but with the additional claim that

the decisions that are most relevant for economics are ones in which people *do* have opportunities to discover and correct anomalous preferences (e.g., Smith, 1989; Harrison, 1994; Binmore, 1999). Plott (1996) presented this argument particularly clearly as the *discovered preference* hypothesis. According to this hypothesis, each individual has coherent preferences, but these preferences are not necessarily revealed in their decisions. Rationality is a 'process of discovery': when individuals face unfamiliar tasks, their behaviour can be influenced by various biases, but with sufficient incentives and after sufficient experience, they arrive at 'considered choices' that reflect stable underlying preferences (p. 248). Thus, 'If principles of psychology have a role to play in explaining social choices, it must be to explain deviations from the general tendencies explained by the rational choice models' (p. 226).

In a comment on Plott's paper, Kahneman (1996) characterises Plott's methodological strategy as retaining the rational choice model wherever it works but invoking special mechanisms to account for anomalies: 'The assumption of rationality is not given up: the rationality of agents is assumed to be latent when it is not manifest in the behaviour of the moment'. Kahneman sees this separation between economic explanation for rational behaviour and psychological explanation for non-rational behaviour as 'deeply problematic' (p. 251). He draws an analogy with the explanation of visual perception:

> Visual perception is remarkably accurate, but it is also prone to systematic illusions and biases. An extension of Plott's strategy would partition the topic of visual perception into the study of accurate perception and the study of errors and illusions. The former would be considered normal and self-explanatory; explanation would only be required for illusions and biases (p. 252).

But, Kahneman says, accurate perception is *not* self-explanatory, and an investigation of biases can identify mechanisms that tend to produce accuracy. He gives the example of over-estimation of distances in fog. This is a bias, but it is also evidence that blur is a cue for judgements of distance. Thus, 'An analysis that takes accurate perception for granted and only explains illusions is not good psychology' (p. 252).

In the late 1990s and early 2000s, Robin Cubitt, Graham Loomes, Chris Starmer and I were running experimental tests of the discovered preference hypothesis and finding that violations of EUT often persisted when participants had repeated incentivised experience – sometimes market experience – of the relevant tasks (Cubitt *et al.*, 2001; Loomes *et al.*, 2003 *et al.*, 2010). We applauded Kahneman's response to Plott as a cogent methodological critique of the hypothesis that was failing our tests.

I still believe that we and Kahneman were right. Nevertheless, Kahneman's argument seems inconsistent with the 'Prospect Theory' concession. The contrast between the careful argument of the 1996 comment and the unsupported assertions of the concession paragraph supports the supposition that that concession may have been merely tactical.

Kahneman and libertarian paternalism

In an informal meeting with Kahneman in (I think) 1999, he expressed enthusiasm for a line of argument that was currently being developed by Sunstein and Thaler and

which later emerged as *libertarian paternalism* (Sunstein and Thaler, 2003). When I read Sunstein and Thaler's paper, I was curious about why Kahneman had liked their argument, since it seemed to me to invoke an implicit assumption of latent rationality. In February 2015, in correspondence with Kahneman about a different matter, I tried to satisfy this curiosity.

With Gerardo Infante and Guilhem Lecouteux, I had co-authored a critique of work in normative behavioural economics that assumed latent rationality (Infante *et al.*, 2016). I sent Kahneman a pre-publication copy of the paper. Here (in abbreviated form) is the resulting email correspondence:

Sugden (05/02/2015): Incidentally, I have recently been working with two co-authors on a critique of what we see as the mainstream approach to behavioural welfare economics. We see this critique as in the same spirit as your 1996 comment (which I have always liked) on Charlie Plott's 'discovered preference hypothesis'. In case you are interested, I attach a copy.

Kahneman (05/02/2015): I look forward to reading your paper. I was disappointed to find that I do not have a copy of my comment on Plott, which you seem to remember better than I do!

Sugden (05/02/2015): I attach a copy of your comment on Plott. The passages I like to quote are in the first five paragraphs, where you say that Plott's strategy assumes 'latent rationality' and treats rationality as something that doesn't need a psychological explanation – psychology is needed only to explain deviations from rationality. Our critique of 'behavioural welfare economics' is that it simply assumes the existence of (coherent) latent preferences and then uses the satisfaction of these preferences as its normative criterion.

Kahneman (07/02/2015): I have read your interesting paper, and you will not be surprised that I disagree with it. I will not speak of Shleifer's approach,[1] but I think I understand Sunstein-Thaler, and believe that your analysis does not accurately represent the position that I share with them.

What we believe (I believe) is that choices are context-dependent, but also that *people have preferences about the context in which they would rather answer a particular type of question*. The psychological claim is that we (people in general) often have the sense that some choices correspond more closely to our 'true self' than others. For example, most people will state that their true self is better expressed when they are 'cool' rather than emotionally aroused, when they have time to reflect, when the questions are couched in broad frames etc. In particular, people who do not currently save may feel that their 'true self' wishes to save, and for many of us our true self eats less sweets than our real self does. Individuals who give inconsistent answers to the 'same' question framed in different ways sometimes know which answer better corresponds to the preferences of their true self. An example is Samuelson's friend,[2] where most

[1]A reference to Infante *et al.*'s discussion of Bordalo *et al.* (2013).
[2]An anomaly in choice under risk discussed by Samuelson (1963).

people (I believe) would say that their answer to the integrated version of the question about how to choose between gambling ten times or none is more representative than their answers to ten questions about successive gambles.

A nice example of how this might work is an informal study of our lost-ticket vs lost-money example.[3] What happens when you ask people both questions in immediate succession? Our informal observation was that when the sequence is lost-ticket → lost money, people often changed their response, from 'not-buy' to 'buy'. But when the sequence was reversed, they stayed with their initial choice (buy new tickets). The implication is that people view their response to the broader frame (money) as more robust and reflective of their 'true self' than their answer to the narrow frame (tickets). I believe that Thaler and Sunstein refer to this 'true self' when they say that libertarian paternalism helps people make the choices they would want to make.

The key point is that the choices of the true self do not demonstrate latent rationality, in the demanding sense of completeness and consistency. There are some choice problems for which the true self has no answer. One example I have worried about is the lives-saved vs lost problem.[4] We observed (again informally) that here people who are confronted with their conflicting choices are dumbfounded. There seems to be no compelling feeling that one of the frames is more revealing of true preferences than the other. Preference is gone, because the moral emotion that the versions of the problem evoke are attached to changes, not to the identical final states. So completeness fails. And I am sure that consistency could be shown to fail as well in the same way: there will be 'true' preferences that conflict – leading to dumbfounding when the individual is confronted with the inconsistency.

This treatment is psychological all the way down. Its normative appeal comes from individuals' own preferences among their conflicting answers – and the treatment offers no solution when the individuals cannot choose (are dumbfounded, a term I borrow from Jonathan Haidt). I do not think that your analysis of the normative content of behavioral economics represents it. The position I share (I believe) with Thaler and Sunstein is quite different from the position taken by Shleifer, Bordalo, and by Matthew Rabin in his more recent work. Like you, these economists are theory-driven, and they necessarily invoke consistency to avoid theoretical vacuity and nonsense. Thaler, Sunstein and I are non-theorists: we consider incoherence a fact of life, are comfortable with the idea that normative appeal comes in degrees, and deny that irreducible inconsistency in a set of choices means that anything goes.

I am not sure any of this makes sense – but I am quite sure that my critique of Plott does not apply to Thaler and Sunstein, and that their claim to speak for individuals' true self has some merit (although caution is advised.) Another way of saying this is that your analysis attributes to us a greater concern with consistency than we actually feel. You probably cannot help doing so, because you see inconsistency as more devastating than we do!

Sugden (07/02/2015): I *was* surprised that you disagreed so strongly, since it still seems to me that our argument is a natural development of the ideas you expressed in your comment on Plott.

[3]An anomaly in riskless choice discussed by Kahneman and Tversky (1982).
[4]The 'Asian disease' problem discussed by Tversky and Kahneman (1981).

I'm not convinced that there is as big a difference as you think between your version of Sunstein-Thaler and ours. In the examples in which you think the Sunstein-Thaler approach works (e.g. savings, sweet-eating) the choices of the real human being are context-dependent, but (on your account) the 'true self' endorses the choices made in one context and disowns those made in the other. So, in these examples, the true self *is* latently rational (i.e. is expressing context-independent preferences rather than the context-dependent preferences of the real human being), and the problem (as faced by the libertarian paternalist) of judging the individual's best interests is being resolved by appeal to latent rationality. We are not claiming that Sunstein and Thaler have to assume that the true self has consistent preferences over *absolutely everything* (as in the traditional neoclassical model) – only that, when S&T confront specific cases of context-dependent preferences, they respond by attributing consistent latent preferences to individuals.

It seems to me that your argument is that in many significant cases (but not all, as in your examples of 'dumbfoundedness'), individuals *really are* latently rational, and that in 'cool' states they can report the preferences of their true selves. I think we probably disagree about how much people think in terms of 'true selves', and (when they do think in this way) how consistently they think of their 'cool' selves as the true ones, rather than being context-dependent in their views about their true selves.

You are perhaps right to say that, in relation to the issues addressed by normative economics and political philosophy (as opposed to those addressed by empirical economics), you and Thaler are 'non-theorists'. (I'd be surprised if Sunstein was happy to be thought a non-theorist in this domain.) But I'm not sure I recognise your characterisation of me as being 'theory driven' and as seeing inconsistency as 'devastating'. It depends what you mean by inconsistency. I don't find it at all devastating to think that inconsistency *in people's preferences and choices* is a fact of life. But yes, I am trying to construct a *theoretically* and *philosophically* consistent way of doing normative economics which can accommodate that fact of life.

Kahneman (07/02/2015): The key point I stressed, which I did not recognize in your paper, is that the selection of an optimal choice context is made by the individual himself. This is a point that Thaler and Sunstein emphasize, and my impression was that you were rather dismissive of it. Furthermore, the selection is ad hoc and specific to a particular choice problem – they have not promoted the idea for more general choices. I do not think that this is quite the same as attributing latent preferences. Is it?

Sugden (08/02/2015): I have read Thaler and Sunstein's position on the cafeteria example and similar cases as saying that the choice architect's criterion should be the preferences that (the choice architect believes) the individual *would have* revealed in the choice context that he *would have* chosen as most representative of his true self. *If* it is an empirical fact that individuals have stable and predictable judgements about which choice contexts are most representative of their true selves, then T&S can indeed say that, in the ideal cafeteria, 'the selection of an optimal choice context is made by the individual himself'. But the 'if …' clause is the assumption that

(in relation to the specific problem under discussion) individuals have latent preferences.

Kahnemam (08/02/2015): Perhaps I now see the issue. As I read your paper (and my comment on Plott) latent preferences are assumed to be rational. I thought you shared this assumption of rationality, but if I now understand you correctly I was wrong. The latent preferences of TS are certainly not rational in the classic sense – only less unreasonable than some alternatives.

Sugden (08/02/2015): Yes, I think we have converged on where we agree and disagree.

Conclusion

What did Kahneman and I converge on? On my reading of this correspondence, Kahneman is retracting the concession paragraph of 'Prospect Theory'. The implication of that paragraph is that individuals have latent preferences that have the rationality properties of EUT. The Kahneman of 2015 endorses his 1996 response to Plott and interprets that response as a criticism of Plott's assumption of *rational* latent preferences.

At the same time, Kahneman is using a concept of *true* preference, defined indirectly in terms of the individual's judgements about the contexts in which their choices would best express their true self. In his second use of the term 'dumbfounding', he effectively adds a further condition – that judgements between contexts are independent of the context in which the judgements themselves are expressed. If, for a given choice set, the individual can identify a uniquely self-expressing context for making the choice (or a set of such contexts which deliver the same decision), his or her true preferences conditional on that choice set are by definition context-independent. Thus, Kahneman's concept of true preference *does* incorporate principles of latent rationality, namely two tiers of context independence. But since true preferences need not be complete, Kahneman is not treating rationality as self-explanatory, as we both think Plott is doing.

However, libertarian paternalism can give policy guidance only when revealed preferences are context-dependent (i.e., when choices can be affected by changes in choice architecture), but well-defined true preferences exist. Kahneman clearly believes that, given his conception of true preferences, this combination is sufficiently common for Sunstein and Thaler's approach to have wide application. I am more sceptical.

Although we both recognise that this disagreement is in large part empirical, I sense that something more is at stake. It is perhaps significant that Kahneman uses 'dumbfounded' to describe the experience of discovering your choices to be context-dependent and of being unable to answer the question of which context delivers the preference of your true self. Dumbfoundedness is an emotion that goes with discovering the inexplicable to be true. It is as if, for Kahneman, you really have a true self which *expects* to have context-independent preferences and finds their absence inexplicable. For me, it is the idea of having a true self that is ultimately problematic.

References

Binmore, K. (1999), 'Why experiment in economics?', *Economic Journal*, **109**: F16–F24.
Bordalo, P., N. Gennaioli and A. Shleifer (2013), 'Salience and consumer choice', *Journal of Political Economy*, **121**: 803–843.
Cubitt, R., C. Starmer and R. Sugden (2001), 'Discovered preferences and the experimental evidence of violations of expected utility theory', *Journal of Economic Methodology*, **8**: 385–414.
Harrison, G. (1994), 'Expected utility and the experimentalists', *Empirical Economics*, **19**: 223–253.
Infante, G., G. Lecouteux and R. Sugden (2016), 'Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics', *Journal of Economic Methodology*, **23**: 1–25.
Kahneman, D. (1996), 'Comment', in K. Arrow, E. Colombatto, M. Perlman and C. Schmidt (eds), *The Rational Foundations of Economic Behaviour*, Basingstoke: Macmillan and International Economic Association, 251–254.
Kahneman, D. and A. Tversky (1979), 'Prospect theory: an analysis of decision under risk', *Econometrica*, **47**: 263–291.
Kahneman, D. and A. Tversky (1982), 'The psychology of preferences', *Scientific American*, **246**(1): 160–173.
Loomes, G., C. Starmer and R. Sugden (2003), 'Do anomalies disappear in repeated markets?', *Economic Journal*, **113**: C153–C166.
Loomes, G., C. Starmer and R. Sugden (2010), 'Preference reversals and disparities between willingness to pay and willingness to accept in repeated markets', *Journal of Economic Psychology*, **31**: 374–387.
Plott, C. (1996), 'Rational Individual Behaviour in Markets and Social Choice Processes: The Discovered Preference Hypothesis', in K. Arrow, E. Colombatto, M. Perlman and C. Schmidt (eds), *The Rational Foundations of Economic Behaviour*, Basingstoke: Macmillan and International Economic Association, 225–250.
Samuelson, P. (1963), 'Risk and uncertainty: a fallacy of large numbers', *Scientia*, **98**: 108–113.
Smith, V. (1989), 'Theory, experiment and economics', *Journal of Economic Perspectives*, **3**: 151–169.
Sunstein, C. and R. Thaler (2003), 'Libertarian paternalism is not an oxymoron', *University of Chicago Law Review*, **70**: 1159–1202.
Tversky, A. and D. Kahneman (1981), 'The framing of decisions and the psychology of choice', *Science*, **211**(4481): 453–458.

Cite this article: Sugden, R. (2025), 'Daniel Kahneman and the concept of the true self', *Behavioural Public Policy* **9**, 285–292. https://doi.org/10.1017/bpp.2024.39