What Daniel Kahneman thought about economics:

a reconstruction and critique

Chris Starmer and Robert Sugden

10 June 2025

Abstract

Drawing on Kahneman's writings, his responses in interviews and our personal recollections, we reconstruct and critically assess his view of his contributions to behavioural economics, of developments in economics that have built on those contributions, and of earlier related work by economists. Kahneman always insisted that he was a psychologist, pure and simple. He saw his primary research programme as the study of intuitive judgement. While he believed that outputs from this programme could be usefully applied in economics, he differentiated his programme from most work in behavioural economics, including that of economists such as Allais and Markowitz who had proposed alternatives to expected utility theory from the 1950s. He saw behavioural economics as distinct from his approach because it retained the fundamental maximising structure of rational choice theory and represented psychological effects in 'one deviation at a time' models. We argue that the two programmes have more in common than Kahneman thought, and that their respective methodologies are complementary and mutually informative. As used by many behaviourally-inclined economists, utility maximisation is a flexible modelling strategy for representing considerations that might enter a reasonable person's decision making, rather than a commitment to the belief that individuals act on preferences that satisfy a priori axioms of normative rationality. The research method of considering one deviation at a time is used in both modelling and experimentation and in both economics and psychology, in ways that promote cumulative advances in scientific knowledge.

(239 words)

1. Introduction: Kahneman's Nobel Lecture

In the aftermath of Daniel Kahneman's death in March 2024, much has been said and written about his role as a founder (some would say *the* founder) of behavioural economics. Naturally enough, the focus has been on Kahneman's work as viewed in the perspective of current behavioural economics. In contrast, our paper focuses on Kahneman's own view of his contributions to behavioural economics, of developments in economics that have built on those contributions, and of related preceding work by economists. Drawing on Kahneman's writings, on his responses to interview questions and on our recollections of personal engagements with him over many years, we will suggest that there are significant divergences between the two views. Awareness of these divergences allows a more rounded understanding of Kahneman's work.

Since we are writing about the relationship between Kahneman's work and 'behavioural economics', we need to explain how we use this term. On our understanding, behavioural economics is the application of concepts and methods of psychology to problems in economics. The term came into general use in the period around 2000, together with the idea that it identifies a distinct approach to economics. There has been an intermittent flow of ideas from psychology to economics throughout the twentieth century and even before,[1] but Kahneman has had a huge influence on the emergence of behavioural economics as a self-conscious sub-discipline.

A seemingly obvious starting point for our enquiry is the Prize Lecture that Kahneman (2002) gave after being awarded the 2002 Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel. The Prize had been awarded jointly to Kahneman and Vernon Smith as the 'pioneers' of 'psychological and experimental economics', described as 'two distinct, but currently converging, areas: the analysis of human judgment and decision-making by cognitive psychologists, and the empirical testing of predictions from economic theory by experimental economists' (NobelPrize.org, 2002). Specifically, Kahneman's Prize was 'for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty'. In the community of behavioural and experimental economists,[2] the Prize was seen as a crowning recognition of the scientific status of their work. It would have been natural for economists to expect Kahneman's Prize Lecture to showcase how psychological research was contributing to their discipline.

In fact, however, the Lecture makes very little reference to economics. Its title is 'Maps of bounded rationality: a perspective on intuitive judgment and choice'. (In effect,

---

[1] For discussions of this flow, see e.g., Sent (2004) and Bruni and Sugden (2007).

[2] In 2002, the term 'behavioural' was just beginning to be used for what the Nobel citation referred to as 'psychological' economics. Many practitioners (including ourselves) preferred to treat Kahneman's and Smith's work as belonging to a single, broadly-defined research programme of 'experimental' economics.

Kahneman treats intuitive choices as judgements about what to choose. From now on, we will use the term 'judgement' to include choice.) The opening section of the Lecture distinguishes between *intuitive* ('*System 1*') and *deliberate* ('*System 2*') thought processes. Kahneman's primary concern is with intuitive judgements—judgements that come to mind spontaneously, as perceptions do, without the intervention of conscious deliberation. The theme of the lecture is the importance of *accessibility* in the automatic processes that create intuitive judgements. Thoughts are accessible to the extent that they come to mind easily. Because accessibility depends on context, intuitive judgements are necessarily context-dependent (2002: 452–456).

The five succeeding sections of the Lecture summarise experimental findings, mainly by psychologists, about different forms of accessibility. The work that most economists would have seen as Kahneman's outstanding contribution (jointly with the then late Amos Tversky) to their discipline features in just one of these sections, 'Changes and states: prospect theory'. The central idea in this section is that perceptions are reference-dependent: perceptual systems 'are designed to enhance the accessibility of changes and differences' rather than to record constant states (2002: 459–460). Kahneman's opening example is about visual perception and the reference-dependence of perceptions of brightness. Prospect theory is then presented informally, as organising and explaining experimental observations of reference dependence in choices between lotteries (2002: 460–465). In his final summing-up, Kahneman re-emphasises the generality of the ideas he has been presenting—the analogy between intuition and perception, the fundamental differences between intuitive thought and deliberate reasoning, and the role of accessibility in intuitive judgement. Their generality is that they 'play a fundamental role in several domains of social and cognitive psychology' (2002: 483). There is no mention of their role in economics.

Unquestionably, the Lecture describes a major research programme in psychology, and one to which Kahneman had made fundamental contributions. Behavioural economists would have understood that this work was important to economics, but probably few other economists—and still less a general audience—would have shared that understanding. It is surely surprising that the first (and so far, the only) psychologist to win the economics Nobel Prize chose not to say more about the significance of his discipline *for economics*.

Each year, it is customary for the *American Economic Review* to publish the preceding year's Prize Lecture(s), sometimes with authorial revisions. A revised version of Kahneman's Lecture was published in the December 2003 issue (Kahneman, 2003b). Kahneman made substantial alterations to his original Lecture. The subtitle was changed from 'A perspective on intuitive judgment and choice' to 'Psychology for behavioral economics', with the implication that the revised version was addressed to economists rather than to a general audience. The 'for' in the subtitle hints at a one-directional relationship between the two disciplines: Kahneman will explain how the findings of a psychological research programme can inform economics.

In the revised introduction, Kahneman notes that economists often criticise psychology for its 'failure to offer a coherent alternative to the rational-agent model'. He acknowledges that 'psychological theories of intuitive thinking cannot match the elegance

and precision of formal normative models of belief and choice', but for Kahneman 'this is just another way of saying that rational models are psychologically unrealistic'. They are unrealistic because, as a matter of empirical fact, 'most judgments and most choices are made intuitively' (2003b: 1449–1450). If economics is to be an empirical science, it has to take account of the psychology of intuitive judgement.

The revised version retains the main content of the Lecture, but adds a lot of additional material about work in behavioural economics from the 1980s. This material serves two main purposes in Kahneman's argument. First, it provides evidence that behaviour *in domains that economic theory claims to explain* shows regularities that are inconsistent with standard theories of rational choice, but are explicable by psychological theories of intuitive judgement. Second, the fact that many of these findings have been generated *by economists* shows that ideas from psychology are now (i.e., by 2003) being applied in economics, with scientifically valuable results. Kahneman allows himself and Tversky some credit for this, but at one remove. In the introduction, he says that he and Tversky were drawn in to a cross-disciplinary conversation by economists (Richard Thaler is named specifically) 'who hoped that psychology could be a useful source of assumptions for economic theorizing, and indirectly a source of hypotheses for economic research', and that these hopes have 'been realised to some extent, giving rise to an active program of research by behavioral economists' (2003b: 1449). If only 'to some extent', the revised version of the Lecture allows behavioural economists to view Kahneman's Nobel Prize as an honour for their research programme.

Nevertheless, Kahneman distances himself from that programme. In the final section of the revised Lecture, he characterises the rational agent of economic theory as 'endowed with a single cognitive system that has the logical ability of a flawless System 2 and the low computing costs of System 1'. He says that theories in behavioural economics have 'generally retained the basic architecture of the rational model, adding assumptions about cognitive limitations designed to account for specific anomalies'. In contrast, in the approach that he has presented in the Lecture, 'the central characteristic of agents is not that they reason poorly but that they often act intuitively. And the behavior of these agents is not guided by what they are able to compute, but by what they happen to see at a given moment' (2003b: 1469). Despite these reservations, however, the revised Lecture ends diplomatically, recognising that behavioural economics is at least moving in the right direction:

> Incorporating a common sense psychology of the intuitive agent into economic models will present difficult challenges, especially for formal theorists. It is encouraging to note, however, that the challenge of incorporating the first wave of psychological findings into economics appeared even more daunting 20 years ago, and that challenge has been met with considerable success. (2003b: 1470)

Shortly after the award of the Nobel Prize, Kahneman expressed similar reservations about behavioural economics in a short paper, 'A Psychological Perspective on Economics', presented at a session on 'Views of economics from neighbouring social sciences' at the annual conference of the American Economic Association in January 2003. The premise of the paper is that, when Kahneman first encountered economic theories of human behaviour in

the early 1970s, there was a wide gap between the assumptions made by economics and psychology. The agents assumed by economics were rational and selfish, and acted on fixed tastes; as a psychologist, Kahneman had been trained 'not to believe a word of [this]'. The paper asks how far this gap had narrowed since that time. After a wide-ranging review of developments in behavioural economics over that period, Kahneman concludes:

> Much has happened in the conversation between economics and psychology over the last 25 years. The church of economics has admitted and even rewarded some scholars who would have been considered heretics in earlier periods, and conventional economic analysis is now being done with assumptions that are often much more psychologically plausible than was true in the past. However, the analytical methodology of economics is stable, and it will inevitably constrain the rapprochement between the disciplines. […] Thus, it now appears likely that the gap between the views in the two disciplines has been permanently narrowed, but there are no immediate prospects of economics and psychology sharing a common theory of human behavior. (2003a: 165–166)

In concluding that rapprochement is constrained by the analytical methodology of economics, Kahneman is referring (as he also does in the revised Lecture) to what he sees as two fundamental features of that methodology—its commitment to a 'basic framework' that derives from a normative conception of rational choice, and its 'noncumulative' strategy of representing deviations from rationality one at a time in otherwise conventional models (2003a: 163, 166).

The implication of the Nobel citation was that, as a result of Kahneman's work, insights from psychological research were being integrated into economic science. It seems that Kahneman believed that some of the most important insights of this work were *not* being integrated into economics. If this is what he thought, tact may have led him to avoid explicit references to economics in the Prize Lecture, and simply to explain the insights themselves.

Our aim in this paper is to clarify how and why Kahneman differentiated his research programme from that (or perhaps those) of behavioural economics, and to assess how well those differentiations reflected the reality on the ground. We begin by looking at Kahneman's accounts of his position on these issues, avoiding critical commentary as far as possible until we have presented the position in its entirety. In Section 2, we review Kahneman's interpretation of his own research programme of investigating intuitive judgements. In Section 3, we examine how he differentiated his and Tversky's approach to decision theory from earlier and concurrent attempts by economists to explain observed tendencies for choices to contravene expected utility theory. In Section 4, we examine Kahneman's reservations about developments in behavioural economics from the 1980s—developments that were often presented by their authors as building on Kahneman and Tversky's contributions. In Section 5, we examine what Kahneman thought about prospect theory itself. We have left this until last because Kahneman seems to have had mixed feelings about prospect theory, and particularly about his and Tversky's later refinements of it—refinements that brought it methodologically closer to economics.

Having set out Kahneman's analysis of the methodological gap between behavioural economics and the psychology of intuitive judgement, we offer our own assessment of its validity. We focus on the two features that, for Kahneman, characterise the 'analytical methodology of economics'. In Section 6, we ask how far economists' empirical theories of human behaviour are, or have been, constrained by predetermined normative assumptions about human rationality. We argue that behavioural economics builds on previous lines of research in which economists interpreted the concept of rationality more flexibly than Kahneman's characterisation of economic methodology would allow. In Section 7, we agree with Kahneman that many models in behavioural economics have a 'one deviation at a time' structure, but argue that this methodological strategy can produce cumulative progress in understanding human behaviour. In the final section, we conclude that Kahneman's psychological methodology of investigating intuitive judgements is complementary with an economics-based methodology in which there is a default assumption of utility maximisation.

2. Intuitive judgement: Kahneman's research programme

Despite having won the highest honour in economics, Kahneman always insisted that he was not an economist and should not be thought of as one. For example, he was interviewed in 2014 by the methodologist Catherine Herfeld. One of her questions began: 'You are often considered to be the founding father of behavioral economics'. Kahneman's reply began: 'First of all, let me stress that I do not consider myself as the originator of behavioral economics. If anybody is the founding father, then it is Richard Thaler. I do not even consider myself to be an economist at all' (Herfeld, forthcoming).

Nor, despite being a co-author of one of the all-time most highly cited papers in decision theory,[3] did Kahneman consider himself a decision theorist. He saw himself as a psychologist pure and simple, a specialist in the study of intuitive thinking. His sensitivity to his study material gave him deep insights into how and where decision theories *failed*, but he did not claim to have the skills (or indeed the desire) to carry out the necessary reconstruction. When Herfeld asked how important he had considered 'formal decision theory' to be when working on prospect theory, his reply was:

> Amos [Tversky] was a decision theorist. He had been doing formal theory; that was what he was doing. So for him, that came completely natural. For me, it didn't. And so, that was very useful, because Amos had just an enormous respect for utility theory, much more than me. On the one hand, that I did not have this immense respect was very helpful, because I could see problems that he couldn't see. So that's the way we were a very good combination. On the other hand, I could not do the theory, so I had nothing to do with the axioms. (Herfeld, forthcoming.)

The Prize Lecture may not have been what behavioural economists expected, but it was a beautifully clear and accessible account of the psychological research programme that

[3] Among papers *in the whole of economics* published in the period 1970–2002 and cited up to 2006, Kahneman and Tversky's (1979) 'Prospect theory' paper ranked as no. 2. See Kim et al. (2006).

Kahneman identified with, and of what it had achieved. That programme is the study of judgements that come to mind automatically in the presence of activating cues.

In some cases, the mental operations that create these judgements are genetically hard-wired. That is probably true of some of the properties of visual perception that Kahneman discusses—for example, the reference dependence of judgements of brightness, and perhaps (one of his favourite examples) the effect of blur on judgements of distance, which induces over-estimation of distance in fog (2002: 465, 471). In other cases, such as the instantaneous decision making that is an essential part of the ability to drive a car, intuitive judgements are the product of experience and training.

Intuitive judgements are closely related to perceptions and emotions. In Kahneman's early work, the main analogy was with perceptions. He credits Paul Slovic with recognising the role of emotions in the construction of intuitive judgements, and refers to later work of his own that used Slovic's *affect heuristic*[4] to explain responses to survey questions about willingness to pay for public goods (Kahneman, 2002: 470; 2003b: 1463). It is inherent to the concept of intuition that intuitive judgements are not direct products of reasoning. Thus, they are not well explained by theories of reasoning—even theories that take account of cognitive limitations. However, because they embody the effects of natural selection and experiential learning, intuitive judgements are generally well adapted to the decision problems that individuals normally face (2011: 21–30).

Kahneman's research identifies general regularities in intuitive judgements, such as the forms of accessibility discussed in the Lecture. Given the connections between intuitive judgements, perceptions and emotions, it should not be surprising that the regularities he finds apply across domains which, as viewed from economics, are very disparate. For example, one of these regularities is *attribute substitution*. This occurs when an individual 'assesses a specified *target attribute* of a judgment object by substituting a related *heuristic attribute* that comes more readily to mind'. Thus, 'people who are confronted with a difficult question sometimes answer an easier one instead' (2002: 466, italics in original). This effect is illustrated in the famous 'Linda' experiment in which participants were given a description of Linda (trained in philosophy, deeply concerned about social justice, etc.) and asked to rank a set of statements about her in order of probability. Many participants judged 'Linda is a bank teller and active in the feminist movement' to be more probable than 'Linda is a bank teller'. The answer that follows from the rules of probability requires a little thought, but intuition can immediately supply the idea that 'Linda fits the idea of a "feminist bank teller" better than she fits the stereotype of bank tellers' (2011:180). An economist might think about this observation in relation to probability theory, but Kahneman sees it in relation to general patterns of intuitive thinking that apply far beyond probability judgements.

Kahneman argues that psychology should not be expected to provide a general theory of intuitive judgement, and does not need one. Summing up his account of experimental findings about accessibility, he says:

---

[4] For a general account of the affect heuristic, see Slovic et al. (2007).

> [M]uch is known about the determinants of accessibility, but there is no general theoretical account of accessibility and no prospect of one emerging soon. In the context of research in judgment and decision making, however, the lack of a theory does little damage to the usefulness of the concept. For most purposes, what matters is that empirical generalizations about the determinants of accessibility are widely accepted—and, of course, that there are procedures for testing their validity. (2002: 456)

And, comparing psychology to economics:

> [T]he alternative to simple and precise models is not chaos. Psychology offers integrative concepts and mid-level generalizations, which gain credibility from their ability to explain ostensibly different phenomena in diverse domains. (2003b: 1449)

In Kahneman's view, the value of theory is over-rated—at least by economists, but perhaps more generally. Here is part of his reply to a question from Herfeld about how economics might benefit from work in psychology:

> [F]or me and for psychologists in general, and for some behavioral economists whom I know, science is about making discoveries, it is not about having a theory, it is about facts. It is about things that are observable and true. And theory is, in a way, secondary as an objective. It's beautiful to have a theory, but if you don't and you're making discoveries then you're making progress. (Herfeld, forthcoming.)

Kahneman sees intuitive judgements as belonging to the System 1 component of a dual-processing theory of cognition. Reasoning is deliberate, and belongs to System 2. Kahneman recognises the ability of System 2, *if activated*, to 'override', 'modify' or 'correct' System 1 judgements (2002: 481; 2003b: 1467). However, there is no need for such corrections 'when all goes smoothly, which is most of the time' (2011: 24).

Psychology needs to explain the mental operations of both systems, as Kahneman makes clear when he summarises the Lecture as setting the following 'agenda for research': 'to understand judgment and choice we must study the determinants of high accessibility, the conditions under which System 2 will override or correct System 1, and the rules of these corrective operations' (2002: 481). However, Kahneman's own work within this agenda focuses on System 1 and on the *accessibility* of System 2 thoughts, i.e., the conditions that activate System 2 (2003b: 1467–1469). What those thoughts are is not a main item on his personal research agenda. To put this another way, his research does not aim for a systematic classification of intuitive judgements according to whether System 2, if activated, would endorse or override them.

Kahneman does not assume that System 2 reasoning reliably tracks objective or widely accepted principles of correctness: 'The rules that people apply in deliberate reasoning are sometimes false' (2002: 472). Nevertheless, unambiguous and easily accessible criteria of correctness can serve as research tools for identifying intuitive judgements and testing hypotheses about them (2002: 465). Kahneman's principal examples of intuitive judgements

are *exhibits* or *bottled phenomena*—replicable experimental designs that reliably produce a particular phenomenon.[5] In some cases, that phenomenon is a single judgement that, on reflection, the person who has made it can easily recognise as incorrect—for example, a preference for a stochastically dominated lottery, or the over-estimation of distance in fog. In other cases, the phenomenon is a pair of judgements, neither of which is clearly correct or incorrect, but which can easily be recognised as mutually inconsistent. Thus, Kahneman says that he and Tversky 'restricted framing effects to discrepancies between choice problems that decision makers, upon reflection, consider effectively identical'. Their famous 'Asian disease' task satisfies this criterion because 'observers agree that it would be frivolous to let a superficial detail of formulation [i.e., whether health policies are described in terms of deaths or lives saved] determine a choice that has life and death consequences' (2002: 457).[6]

As we noted in the Introduction, Kahneman argues that intuitive judgements are the source of many observed regularities in economic behaviour that contravene traditional theories of rational choice. Countering an argument made by some defenders of those theories, he cites evidence that violations of rational choice theory do not become less frequent as financial stakes increase. High stakes can be expected to increase the effort and attention that a person applies to a decision task, but where that extra effort and attention are applied may still be governed by intuitive mental operations (2003b: 1469).

Nevertheless, Kahneman seems to have accepted that the axioms of expected utility theory (EUT) are *normatively* justified. One of Herfeld's questions was: '[I]n your work you considered normative expected utility theory useful as offering a standard. Is that something you would still argue for?' Kahneman replied:

> [Tversky and I] never changed our mind, really, about the normative appeal of utility theory. […] I've never been terribly interested in normative theory, I am not a decision theorist. Of the two of us, Amos Tversky was the decision theorist. But we never saw any point in changing the traditional theory. It looks perfectly sensible and the requirements of consistency appear quite sensible, well, it's completely impossible psychologically, but it's better if it's satisfied, we never questioned that (Herfeld, forthcoming.)

As reported by Kahneman, Tversky's attitude to rationality was one of *methodological dualism*: he believed that enquiry into the true principles of rationality was a

---

[5] The role of exhibits in experimental economics is discussed by Sugden (2005) and Bardsley et al. (2010: 141–195). In informal conversations with us, Kahneman has called these designs 'bottled phenomena'. We have recently discovered the source of this term. In *A Conversation with Gary Klein* in 2011, Kahenman says: 'Many of us in cognitive and social psychology are engaged in the important exercise that [the Stanford psychologist] Lee Ross has wonderfully described as "bottling phenomena" and our theories are built to fit what we bottle' (https://www.edge.org/conversation/gary_klein-insight, accessed 9 June 2025).

[6] These two types of exhibit have different implications for System 2. By assumption, System 2 has access to modes of reasoning that can *correct the error* in a single-judgement exhibit and can *recognise the inconsistency* in a paired-judgement exhibit. But it does not necessarily have the resources to determine which (if either) of the paired judgements is correct.

meaningful and valuable pursuit, but one that had no necessary connection with empirical investigations of actual human judgements and choices. As a *normative* decision theorist, Tversky endorsed the axioms of EUT. When Tversky first introduced Kahneman to Neumann and Morgenstern's derivation of EUT, 'he presented it as an object of awe' (Kahneman, 2011: 312). The tone and context of this remark suggest that Kahneman recognised Tversky's feelings of admiration for simple, precise and elegant theories, but did not altogether share them: he was more interested in what was observable and true. It seems that Kahneman went along with Tversky's methodological stance on normative rationality, but without any strong commitment: he saw it as orthogonal to his own research programme. Thus, when proposing prospect theory, Kahneman and Tversky (1979: 277) were able to write that its predicted violations of EUT would lead to 'normatively unacceptable consequences'.[7]

3. Antecedents of behavioural economics, as viewed by Kahneman

Kahneman consistently maintained that his and Tversky's development of prospect theory marked a fundamental break with previous economic theory—including the work of economists who had discovered regularities in human behaviour that violated EUT and had proposed alternative decision theories to explain those observations.

In his 'psychological perspective on economics', Kahneman notes that 'the standard of rationality in economics [in the 1970s] was, and [in 2003] remains, the maximization of subjective expected utility'. However, he acknowledges strands of economic literature from the 1950s onwards that challenged this definition of rationality. He refers specifically to Maurice Allais's (1953) and Daniel Ellsberg's (1961) demonstrations of preferences that violate EUT but have 'considerable normative appeal', and to Herbert Simon's (1955) concepts of satisficing and bounded rationality.

Differentiating his work from this literature and citing his and Tversky's work on framing effects in the 1980s (Tversky and Kahneman, 1986), Kahneman says that he and Tversky:

> articulated a direct challenge to the rationality assumption itself, based on experimental demonstrations in which preferences were affected predictably by the framing of decision problems, or by the procedure used to elicit preferences… Unlike the paradoxes of expected-utility theory, [such effects] cannot be defended as normative. (2003a:163).

In this passage, Kahneman does not claim that his and Tversky's explanations *of the paradoxes of EUT* challenge the standard assumption of human rationality: the challenge

[7] This normative statement is followed by an *empirical* concession to advocates of EUT—that these 'anomalies of preference are normally corrected by the decision maker when he realizes that his preferences are inconsistent, intransitive, or inadmissible'. Thus, the anomalies occur only if 'the decision maker does not have the opportunity to discover that his preferences could violate decision rules that he wishes to obey'. However, this concession is retracted when Tversky and Kahneman (1992: 317) propose the 'cumulative' version of prospect theory. We conjecture that it had been a response to a critical referee. For more on this, see Sugden (2024).

comes from their findings about framing effects. But when looking back on his career in *Thinking, Fast and Slow*, he does make that claim. He says that when he and Tversky began their joint work:

> [O]ne of our initial goals was to develop a satisfactory psychological account of Allais's paradox. Most decision theorists, notably including Allais, maintained their belief in human rationality and tried to bend the rules of rational choice to make the Allais pattern permissible. Over the years there have been multiple efforts to find a plausible justification for the certainty effect, none very convincing. Amos had little patience for these efforts; he called the theorists who tried to rationalize violations of utility theory 'lawyers for the misguided'. We went in another direction. We retained utility theory as a logic of rational choice but abandoned the idea that people are perfectly rational choosers. We took on the task of developing a psychological theory that would describe the choices people make, regardless of whether they are rational. (Kahneman, 2011: 314).

Kahneman describes this work as a five-year project that culminated in the 1979 'Prospect Theory' paper:

> Our theory was closely modeled on utility theory but departed from it in fundamental ways. Most important, our model was purely descriptive, and its goal was to document and explain systematic violations of the axioms on rationality in choices between gambles. (2011: 271)

That theory differs from EUT in two key respects. First, utility (or *value*) is defined as a function of *changes in*, rather than *levels of*, wealth. In line with the psychophysical principle of diminishing sensitivity, this function is concave for gains and convex for losses. It is also loss averse, i.e., the disvalue of a loss is greater in absolute terms than the value of an equal and opposite gain. Second, a *weighting function* assigns a *decision weight* to each numerical probability. In the 1979 version of prospect theory, this function satisfies a long list of assumptions which Tversky and Kahneman (1992) later simplified to the 'inverse S' functional form.[8] The assumption that explains the Allais paradoxes[9] is *subcertainty*, which 'captures an essential element of people's attitudes to uncertain events, namely that the sum of the weights associated with complementary events is typically less than the weight associated with the certain event' (Kahneman and Tversky, 1979: 282). As Kahneman and Tversky (1979: 276–277) acknowledge, each of the main properties of their theory had previously been proposed in some form, sometimes by psychologists and sometimes by economists, but the combination was new. The theory was able to explain not only the Allais

[8] The 1979 function incorporates assumptions of subcertainty, subproportionality, subadditivity for small probabilities, overweighting for small probabilities, and discontinuities at probabilities close to 0 and 1.

[9] Kahneman and Tversky (1979) replicate and propose explanations for two exhibits first found by Allais—the *common consequence effect* (Problems 1 and 2 in Kahneman and Tversky's paper) and the *common ratio effect* (Problems 3 and 4).

paradoxes but also the new and surprising *reflection effect*—that reversing the signs of the payoffs in the Allais problems reversed the direction of the violation of EUT.

Describing how he and Tversky arrived at this theory, Kahneman (2011: 271) recounts how the two of them spent many hours examining their own intuitive preferences, 'inventing simple decision problems and asking ourselves how we would choose'. But that is not the whole story. Recall that their starting point was an attempt to explain known regularities in *other people's* behaviour in the simple decision problems that *Allais* had invented. In Kahneman's account of how he and Tversky 'stumbled on' what he calls the 'central flaw' of EUT, Kahneman had been puzzled by the fact that decision theories defined utility as a function of levels of, rather than changes in, wealth. He asked Tversky, as an expert in decision theory, to explain why. Tversky 'remembered that the economist Harry Markowitz … had proposed a theory in which utilities were attached to changes in wealth'. Thinking about Markowitz's theory, they 'quickly concluded that this was the way to go' (Kahneman, 2011: 278–297). Clearly, Markowitz (1952) and Allais (1953) made important contributions to the development of prospect theory. Our interpretation is that Kahneman classified Markowitz's work, like Allais's, as an attempt to maintain the fundamental assumption of human rationality by ad hoc bending of normatively valid rules of rational choice.

Kahneman's desire to distance his own work from what he saw as revised forms of rational choice theory became clear to us (Starmer and Sugden) in a conversation we had with him in 2004. Working with Graham Loomes on choice under uncertainty in the 1980s, we had seen ourselves as members of a small but growing group of economists who, in parallel with Kahneman, Tversky and other psychologists such as Sarah Lichtenstein and Slovic (1971), were discovering regularities in individual decision-making and developing theories to explain them. We had viewed this research programme as including, for example, the work of the economists Jagdish Handa (1977), Ole Hagen (1979), Soo Hong Chew and Kenneth MacCrimmon (1979), David Bell (1982), Mark Machina (1982) and John Quiggin (1982). When 'behavioural economics' began to be promoted as a distinct approach to economics, we were surprised that none of this work—nor indeed that of Allais—seemed to be classified as behavioural. We asked Kahneman what he thought about this exclusion. His response was that this research programme in economics—in which he included our work on regret theory—was based on assumptions about rationality and therefore not 'behavioural'.

4. Kahneman's reservations about behavioural economics

In his 'psychological perspective on economics', Kahneman (2003a) acknowledged that by 2003 many economists were aware of the evidence of violations of rational choice theories, and that this had 'legitimized and encouraged the development of economic theories that model departures from economic rationality in specific contexts'—for example, David Laibson's (1997) model of quasi-hyperbolic discounting. However, Kahneman argued that 'the rationality model continues to provide the basic framework even for these models, in which the agents are "fully rational, except for ..." some particular deviation that explains a family of anomalies' (Kahneman, 2003a: 163). For short, we will call such models *one-*

deviation models. ('One deviation' should not be read literally: models that include only a small number of deviations will be included in this category.)

Kahneman expressed the same reservation about behavioural economics a decade later, in the Herfeld interview:

> My problem is that in most of the papers [by behavioural economists] that I see, they study one anomaly at a time, and it's not that out of all the anomalies that they have studied, behavioral economists have a model of the human being and then they add something to it. They always start with complete rationality and with one exception and then one follows that exception. Rationality plays an essential role in modern behavioral economics, so far as I can see, in the theory. (Herfeld, forthcoming.)

For Kahneman, this one-deviation methodology has two fundamental limitations: it is unable to make a fundamental shift away from the rational choice model of human behaviour, even if such a shift would lead to scientific advances, and it does not allow *cumulative* progress:

> One consequence is that the models of behavioral economics cannot stray too far from the original set of assumptions. Another consequence is that theoretical innovations in behavioral economics may be destined to be noncumulative: when a new model is developed to account for an anomaly of the basic theory, the parameters that were modified in earlier models will often be restored to their original settings. (2003a: 166)

Kahneman recognises that economics needs tractable models, and accepts that the one-deviation methodology may have a pragmatic justification *in economics*. He is not an economist, and does not presume to tell economists what their methodology should be (2003a: 166; 2011: 286–287). But the objective of *his* research programme in psychology is to contribute to the development of an integrated theory *of the human being*. The implication is that if behavioural economics aspires to contribute to this programme, its methodology has fundamental limitations.

Notice again the underlying suggestion of a one-way relationship between psychology and economics. Psychology can discover general properties of human judgement that are liable to have significant effects in the domain of economics. By being aware of these effects and by incorporating them into simple tractable models, economists can arrive at better explanations of phenomena in that domain. But there is no suggestion that economists who work in this way can contribute to the research programme *of psychology*.

5. Prospect theory, as viewed by Kahneman

Kahneman's one-directional view of the relationship between psychology and economics does not fit easily with his account of how he and Tversky drew on the work of Allais and Markowitz in developing prospect theory. Given his reservations about the theoretical approaches characteristic of behavioural economics, it is natural to wonder how far they

might have extended to prospect theory itself. This comment from Kahneman's interview with Herfeld provides some insight:

> There was a field of empirical decision science, and we were working in that field. That field was very strongly influenced by utility theory. […] We played by the rules of what it took to write a good theory within that field. But one argument that I have been making is that what made prospect theory important and what made prospect theory acceptable are two completely different things. What made it acceptable was that it was a pretty good theory in a very limited domain. But that's not why it was important. It was important because of another few ideas, mainly the reference point in loss aversion and those ideas that we used. (Herfeld, forthcoming.)

This comment highlights that, for Kahneman, prospect theory was important primarily as a vehicle for demonstrating the decision-theoretic implications of mid-level psychological generalisations about intuitive judgments—especially loss aversion. Describing prospect theory as 'a pretty good theory in a very limited domain' (i.e., applicable to a restricted domain of lotteries) hints at reservations about the formal modelling approach, but producing such a model was 'playing by the rules' of contemporary economics and decision science (i.e., creating a single formal theory with a utility maximising core) which had the instrumental value, from Kahneman's point of view, of creating a bridge for the flow of ideas from psychology to economics.

In considering Kahneman's attitude to prospect theory, however, it is important to distinguish between two variants of prospect theory: the *original prospect theory* (OPT) set out in Kahneman and Tversky (1979) and the later *cumulative prospect theory* (CPT) presented in Tversky and Kahneman (1992).

OPT is a complex structure presented as a unifying explanation of many regularities in choice under risk, particularly the Allais paradoxes, loss aversion, and risk-seeking in losses. The model, which applies only to the quite narrow domain of simple lotteries (or 'prospects') with no more than two non-zero outcomes, proposes a two-phase theory of choice. In the second 'evaluation phase', prospects are assumed to be ordered by a preference function that is a variant of EUT with the two key modifications explained earlier: the subjective 'value' assigned to any consequence is defined in terms of gains and losses relative to a reference point, and objective probabilities of consequences are transformed to 'decision weights'. While particular forms of non-linear transformation provide the mechanism for explaining the Allais paradoxes in OPT, this approach also entails that the OPT value function predicts violations of stochastic dominance for choices among some sets of prospects, a feature of some significance which we return to below.

Evaluation via this preference function is preceded in OPT by an 'editing phase' during which the set of objectively available options may be modified by the application of various editing operations or heuristics. Most of these can be interpreted as forms of intuitive judgement involving attribute substitution. These include the operations of *simplification* (e.g., rounding probabilities and money amounts), *cancellation of common components* (e.g.,

ignoring the common first stage of a choice between multi-stage lotteries), *combination* (combining the probabilities associated with identical outcomes) and *segregation* (e.g., evaluating a two outcome lottery such as [£30, 0.8; £20, 0.2][10] as the sum of the values of [£20, 1] and [£10, 0.2] to separately evaluate its risky and riskless components). One particular editing operation, however, plays a more specific role in the theory: this is the *dominance heuristic* which eliminates stochastically dominated options, prior to evaluation, if the chooser spots them. With the inclusion of this heuristic, OPT implies that *transparently dominated* options will never be selected, but dominance violation may still arise in cases where dominated options are harder to spot.

While the evaluation phase of OPT can be considered an amalgam of features that had been proposed in previous theories, embedding it in this two-phase structure with editing was a very distinctive approach, relative to any previous models of risky choice proposed in economics, in making explicit reference to cognitive aspects of decision processes. While the resulting theory involves multiple different mechanisms, it coheres with Kahneman's methodology as an application of mid-level generalisations to a specific type of decision, with the editing phase representing the operation of real intuitive judgements assumed to be at work in people's actual choices.

The later CPT model, however, gives much less prominence to the editing phase of decision making (referred to as 'framing' in the 1992 paper). In the opening preamble to the presentation of the theory and in a passage in the concluding discussion, Tversky and Kahneman (1992: 299, 317) nod to the idea that real choices (or at least 'complex' ones) always involve cognitive processes akin to editing/framing to simplify decisions, prior to evaluation. However, CPT is presented as a model of the evaluation phase of decision making only, with no analysis of editing. Moreover, the absence of an explicit editing phase is seemingly downplayed through assertions to the effect that 'The present theory retains the major features of the original version of prospect theory' (Tversky and Kahneman, 1992: 317).

The key new aspect to evaluation in the CPT model is that decision weights, instead of being a simple non-linear transformation of individual probabilities as in OPT, are determined through a 'cumulative' or 'rank-dependent' construction. In this approach, due to Quiggin (1982) and previously adapted to the context of reference-dependent risk preferences by Starmer and Sugden (1989), the weight assigned to any outcome of a prospect depends not only on its own probability, but also its rank position in the prospect's distribution of consequences. Tversky and Kahneman (1992: 299) motivate this more complex approach to probability transformation as a solution to two 'problems' with the original approach:

> This scheme [i.e. simple weighting in OPT, in the absence of editing operations] encounters two problems. First, it does not always satisfy stochastic dominance, an assumption that many theorists are reluctant to give up. Second, it is not readily extended to prospects with a large number of outcomes. These problems can be handled by assuming that transparently dominated prospects are eliminated in the

[10] I.e., a lottery that gives £30 with probability 0.8 and £20 otherwise.

> editing phase [i.e., as in OPT], and by normalizing the weights so that they add to unity. Alternatively, both problems can be solved by the rank-dependent or cumulative functional (Tversky and Kahneman, 1992: 299).

It is doubtful whether Kahneman, if writing as a sole author, would have considered the first 'problem' a problem at all, given that violations of transparent dominance are ruled out by an editing operation of OPT. We conjecture that he may have considered the development of CPT a retrograde step in certain respects. That development jettisons components of OPT that represented the operation of intuitive judgements. As a result, CPT is essentially a composite model of pure utility maximisation using component parts previously developed by economists. Given Kahneman's view that the fundamental innovation of prospect theory was its aim of describing people's actual choices, regardless of their rationality (see Section 3 above), it is hard to see why he would want to treat the satisfaction of a normative principle of rational choice as a selling point for the theory.

This is particularly puzzling, given that the move from OPT to CPT entails a loss of highly distinctive predictive content. Prior to the publication of OPT, Tversky and Kahneman (1986: 263–265) reported extensive violations of stochastic dominance that were turned off or on depending on whether the problem framing renders the dominance relationship transparent or not. Quiggin (1982) had shown that, in the absence of editing operations, OPT predicts violations of dominance, and that if a dominance-elimination operation is included, it predicts violations of transitivity. By using cumulative probability weighting, CPT eliminated these 'problems'. But Starmer (1999) reported experimental evidence of exactly the transitivity violations predicted by OPT and ruled out by CPT.

In a conversation in 1999, we (Starmer and Sugden) told Kahneman about Starmer's experimental result and teased him about the fact that an attempt to make prospect theory more rational had removed its ability to explain a significant effect of intuitive judgement in actual decision making. Far from being disturbed by this, Kahneman seemed both pleased and unsurprised that, in this case, OPT had out-performed CPT. His comment on the removal of the dominance heuristic was 'That was Amos': he (Kahneman) had always liked the editing operations of OPT, and Tversky had been too eager to appeal to economists.

But if Kahneman really did think OPT was a superior theory, what explains his involvement in the development and publication of CPT? A plausible answer is that Tversky was the main driver of the CPT project, aided by researchers who shared enthusiasm for its ambitions.[11] Viewing CPT as ironing out problems in OPT, it would have been natural for Tversky to want to include Kahneman in the project. Kahneman may have been willing to go along with this as a way of promoting key ideas that survived the reformulation (especially loss aversion) to an audience of economists and decision theorists who, in contrast to him, did place a premium on simple and elegant theories that comply with normatively appealing principles. So, notwithstanding the main title of the 1992 paper ('Advances in Prospect

[11] In the acknowledgements to the CPT paper, Tversky and Kahneman express gratitude to 'Peter P. Wakker for his invaluable input and contribution to the axiomatic analysis' and to 'Richard Gonzalez and Amy Hayes for running the experiment and analyzing the data'.

Theory'), we conjecture that Kahneman may have viewed it as something of a retreat from the modelling strategy of OPT—a strategy that was considerably better aligned with his methodological instincts.

6. Is economics committed to a normative theory of rationality?

So far, our aim has been to summarise and understand Kahneman's accounts of the relationship between his research programme and related work in decision theory and economics. These accounts are based on his interpretation of that related work. But, as Kahneman often pointed out, he was neither a decision theorist nor an economist. It is clear from these accounts that his view of decision theory derived in large part from his collaboration with Tversky, and that his view of economics was strongly influenced by Thaler. It seems to us that Kahneman was not attuned to the variety of approaches used in decision theory and economics.

This is particularly evident in Kahneman's treatment of alternatives to EUT that were developed before the publication of 'Prospect Theory'. In Kahneman's account, these theories were fundamentally normative, trying to justify the rationality of observed violations of EUT by making ad hoc and unconvincing changes to the principles of rationality. In this context, Kahneman's conception of normative theory seems similar to that of Leonard Savage (1954). Savage's canonical axiomatization of EUT is intended to provide subjectivist foundations for probability theory. He presents the axioms as principles of mutual consistency among preferences, analogous with principles of logic. In relation to actual decisions, these principles are normative: '[T]he main use I would make of [the axioms] is normative, to police my own decisions for consistency and, where possible, to make complicated decisions depend on simpler ones'. Savage allows that his theory can also be interpreted as 'a crude and shallow empirical theory predicting the behaviour of people making decisions', but thinks that such a theory would be practically useful only in 'suitably limited domains' (1954: 20). On this view, the research programme of decision theory is essentially normative. But, in explaining violations of EUT, neither Allais nor Markowitz located their work in that programme.

Presenting his analysis as a theory of the behaviour of 'the rational man', Allais (1953) distinguishes two different concepts of rationality. The 'abstract' definition requires only that an individual has a preference ordering over probability distributions of outcomes and that this ordering respects stochastic dominance. This is a relaxation, rather than a bending, of the EUT rules of rational choice. Alternatively, rationality can be defined 'experimentally', by 'observing the actions of people who can be regarded as acting in a rational manner'. Roughly speaking, choices made under risk are rational to the extent that they are reasonable and prudent. Allais's criticism of EUT is that its category of irrational behaviour includes actions that are intuitively reasonable and that are predictably taken by people who 'are considered rational by public opinion' (1953: 504–505).

Allais's explanation of his paradoxes combines intuitive psychology with ideas from economics. One of the earliest advances in neoclassical economics was the recognition that,

because consumption goods can be complements of or substitutes for one another, the utility of a bundle of different goods may not be separable into the utilities of its components. Allais's insight was that probabilistic lottery outcomes that combine to make a certainty can be perceived as complements in the same sense (1953: 528n33). It is because EUT does not allow this kind of complementarity that it classifies the 'certainty effect' found by Allais as irrational. In motivating the 'subcertainty' property of the weighting function in OPT, Kahneman and Tversky (1979: 282) use essentially the same explanation of the certainty effect, but represent it as a bias in judgements about actual probabilities (see Section 3 above). A theory that treated such a bias as rational might reasonably be accused of bending the rules of rational choice theory. But this is not Allais's theory. Allais is using a concept of complementarity in preferences that is grounded in economists' experience of explaining human behaviour.

In passing, notice how a theoretical idea (complementarity) developed by economists to explain phenomena in a core domain of their discipline (consumer choice) has had a significant application in psychology. A further application of the same idea can be found in Tversky and Kahneman's (1991) extension of prospect theory to explain reference-dependent preferences over riskless consumption bundles. In their first attempts at this extension, Kahneman and Tversky assumed an additively separable utility function and defined loss aversion separately for each good. This assumption implied a new kind of reflection effect—that indifference curves in commodity space are convex *to the reference point* in the domains of both gains *and losses*. They were surprised to find that this prediction was not confirmed in their early experiments. As shown by Munro and Sugden (2003), the final version of the extended theory can be read as an attempt to maintain separability-based intuitions while allowing complementarity effects.[12]

The starting point for Markowitz's (1952) theory of reference-dependent utility is the problem of explaining the fact that many people take out actuarially unfair insurance against risks of large losses while also engaging in actuarially unfair low-stake gambling. Markowitz demonstrates the implausible implications of Milton Friedman and Savage's (1947) explanation, which uses EUT under the standard assumption that utility is a function of a person's wealth level. He shows that the problem can be resolved by defining utility as a function of changes in wealth relative to a 'customary' level. He presents another piece of evidence of reference-dependent utility – the 'common observation', confirmed in an experiment by Frederick Mosteller and Philip Nogee (1951), that participants in gambling sessions are more willing to take risks after making modest winnings than after incurring modest losses. Markowitz's explanation uses the idea that, as in prospect theory, risk attitudes in the loss domain are the opposite of those in the gains domain. Markowitz also shows that the 'common observation' is consistent with a preference for positive skew in the frequency

[12] In an informal communication with Sugden shortly before Tversky's death, Kahneman reported that Tversky (but, by implication, perhaps not Kahneman himself) had been disturbed by discovering from an early version of Munro and Sugden (2003) that Tversky and Kahneman's (1991) theory allowed cycles of reference-dependent decisions, and had liked the way that Munro and Sugden's axioms ruled out this possibility. There is an interesting parallel with the early development of CPT.

distribution of end-of-session outcomes, anticipating later explanations of Allais's paradoxes. Summing up the achievements of his paper, Markowitz (1952: 158) says that he has 'tried to present, motivate, and, to a certain extent, justify and make plausible a hypothesis which should be kept in mind when explaining phenomena or designing experiments concerning behavior under risk or uncertainty'. Clearly, he sees his work as primarily descriptive rather than normative and thinks it important that empirical hypotheses are psychologically plausible. The 'to a certain extent' qualification to 'justify' suggests that his conception of rationality is one of reasonableness, and that he does not feel the need to take a position on whether reference-dependence is rational or irrational in a more abstract sense.

As an example of later work on alternatives to EUT, we consider the *certainty equivalence* (CE) theory proposed by Handa (1977) shortly before the publication of 'Prospect Theory'. CE is a theory of choice between prospects that are defined in terms of cardinally measurable outcomes (e.g., amounts of money) and their numerical probabilities. In EUT, the subjective value of a prospect is determined by multiplying the probability of each outcome by an amount that is a 'utility' function of the outcome itself. CE is the obverse of this: each outcome is multiplied by an amount that is a function of its probability. That function has essentially the same role as the weighting function in OPT; Handa assumes it to have an inverse-S shape, as in CPT. Handa (1977: 114) does not claim originality for this theory, explaining that it was 'presented in its empirically relevant form' by the experimental psychologists Malcolm Preston and Philip Baratta (1948), and citing robust supporting evidence from experiments conducted in the late 1940s and early 1950s. The novelty of Handa's contribution is an axiomatisation of that theory. In this sense, Handa is an abstract theorist. However, he does not present his axioms as normative principles of rationality. Discussing the crucial axiom that differentiates CE from EUT, Handa says that it 'seems to be quite reasonable and plausible' in relation to 'the prospects that a normal individual generally considers often and seriously' (1977: 103) and that its validity is an empirical matter, 'not in any sense an aspect of "rationality", irrespective of how that term is defined' (1997: 105). He leaves it to the reader to 'form his own opinion on the relative normativeness' of CE and EUT (1997: 115).

These examples challenge the notion that alternatives to EUT, developed by economists prior to prospect theory, were intended to be normative. The models of Allais, Markowitz, and Handa were explicitly descriptive, adapting the EUT framework to capture intuitively reasonable patterns of observed behaviour that contradicted its predictions. A similar interpretation applies to several non-expected utility models developed after prospect theory including, for example, the versions of regret theory and disappointment theory due to Loomes and Sugden (1982, 1986).[13] While formalized as utility-maximization models and sometimes presented axiomatically (e.g., Sugden's (1993) axiomatization of regret theory), these models were not framed as articulations of what rationality requires under risk or

[13] 'Non-expected utility' is a standard term for theories of choice under risk or uncertainty in which individuals act on some kind of utility function but do not always maximise the mathematical expectation of utility.

uncertainty. Instead, they can be interpreted as expanding the scope of considerations that might enter a reasonable person's decision-making.

What unifies these models in behavioural economics—whether developed before or after prospect theory—is their shared attachment to utility maximization as a flexible modelling strategy. In many cases, their developers sought to build better descriptive models by incorporating intuitively plausible motivations absent from the standard model of EUT. At the same time, they often took a neutral stance on whether these models represented normative accounts of rationality in an abstract sense, focusing instead on empirical plausibility and psychological relevance.

Kahneman is right to say that the utility-maximisation framework cannot easily take account of many of the mechanisms of intuitive judgement that his research programme has identified—for example, the probabilistic reasoning anomaly revealed in the 'Linda' problem or the 'Asian disease' framing effect. However, using a model in which some human motivation is represented as utility maximisation is not equivalent to claiming that the behaviour that the model explains is the product of deliberate reasoning. Recall Kahneman's remark that System 2 is not needed 'when all goes smoothly, which is most of the time'. The implication is that intuitively plausible motivations normally produce behaviour through System 1 processes—that is, through intuitive judgements. Models with a utility-maximising structure can often be interpreted as stylised representations of assumed human motivations, without any connotations of actual reasoning processes.

7. One-deviation modelling

While we have questioned Kahneman's interpretation of behavioural economics as fundamentally normative in orientation, we agree with his claim (discussed in Section 4) that much of behavioural economics has pursued a one-deviation modelling strategy of anchoring modelling on the standard framework of rational choice and developing alternatives that introduce one (or at most a few) specific deviation(s) from it. This strategy is typical in the historical development of non-expected utility models (discussed in Section 6). It is common in behavioural economics more broadly, including in the study of time preferences (e.g., Laibson, 1997) and social preferences (e.g., Fehr and Schmidt, 1999). In these literatures, the assumption of utility maximisation is retained and deviations are introduced by introducing new parameters to otherwise conventional utility functions. However, we think that Kahneman's evaluation of this modelling strategy is too negative.

Recall Kahneman's 'fully rational, except for…' description of one-deviation models. It seems to us that he is conflating the use of utility-maximisation as a general modelling strategy with an assumption that each individual's complete set of preferences and beliefs satisfy the kind of consistency requirements that are represented in normative decision theory—the assumption that, in the Herfeld interview, he rightly described as 'completely impossible psychologically'. In many one-deviation exercises in economics, the background model to which the deviation is added is a simplified formalisation of *a*—but not the *only*—psychologically plausible human motivation. For example, if viewed without reference to the

Neumann–Morgenstern or Savage axioms, EUT can be interpreted as a relatively simple representation of two core ideas—that people have preferences over possible outcomes and that they prefer larger probabilities of preferred outcomes to smaller probabilities. In theories of time preference, the background model is one in which, at any point in time, individuals are concerned about the consumption they can expect to enjoy in the current and future periods. In social preference theories, the background model is self-interest.

Kahneman is clearly right that working with a one-deviation modelling strategy cannot lead directly to a comprehensive theory of human decision making. But that is not to say that this research strategy is non-cumulative.

Here it is useful to distinguish between *best buy models* and *models of mechanisms*.[14] Think of a best buy model as the best available model for a specific application, given the current state of knowledge and after making appropriate trade-offs between parsimony and fit to data. By contrast, a model of a mechanism can be thought of as exploring the consequences of introducing a single explanatory factor that might cause deviations from some background model, holding other factors constant.

Because of the value of parsimony, best buy models typically include only a subset of the mechanisms that are known to have some effect on decisions; choosing between alternative models involves judging which mechanisms are most important. The fact that some model is currently in standard use in an empirical science suggests (but certainly does not necessitate) that it has significant best buy properties, with the implication that if a better buy is to be found, it is likely to be a one-deviation model. Thus, one-deviation modelling has a pragmatic justification within the best buy approach. However, that approach has severe limitations as a method of scientific discovery. For example, the development of prospect theory and other non-expected utility models prompted studies which investigated whether any of these new models could qualify as a best buy, together with or supplanting EUT. Although there was no dispute about the ability of many of the new models to explain particular (but often different) systematic violations of EUT, there was reasoned debate about whether EUT remained the best buy (e.g., Harless and Camerer 1994; Hey and Orme, 1994).

A large part of the literature developing and experimentally testing one-deviation models is more appropriately considered as seeking to model and test for the impact of specific mechanisms, without claiming that the mechanism under investigation is the only or most important one in the relevant domain. Examples in this tradition would include our work together with Loomes developing and testing models that incorporate either regret or disappointment, but not both together (Loomes and Sugden, 1982, 1986; Starmer and Sugden, 1989). In the context of social preference, the one-deviation approach has produced models of, for example, reciprocity (Rabin, 1993), inequality aversion (Fehr and Schmidt, 1999) and guilt aversion (Battigalli and Dufwenberg, 2007). We suggest that models of this kind serve functions in the research process that Kahneman does not fully acknowledge in his discussion of behavioural economics.

---

[14] We have discussed this distinction in Sugden (2005) and Bardsley et al. (2009).

The idea that some mechanism affects behaviour *when other factors are held constant* is meaningful only in relation to some benchmark concept that defines constancy. In Kahneman's account of his and Tversky's methodology for investigating intuitive judgements, the benchmark is provided by unambiguous and easily accessible criteria of correctness, such as the principle that judgements about public health policies should be invariant to whether their effects are described in terms of deaths or lives saved (see Section 2 above). Kahneman (2002: 456) describes the framing effects discovered by this methodology as demonstrations of violations of *invariance*—'the assumption that preferences are not affected by variations of irrelevant features of options or outcomes'—and says that invariance is 'an essential aspect of rationality'. In other words, the benchmark is a principle of rationality, and the exhibits that demonstrate framing effects have a one-deviation structure. Tversky and Kahneman are not not assuming that *in the real world* people are 'fully rational, except for' their susceptibility to framing effects. They are constructing pairs of artificially controlled decision tasks that are equivalent according to a benchmark principle of rationality and then investigating whether the two tasks produce systematic differences in individuals' responses.

In the edited version of the Nobel Lecture, Kahneman gives the following overview of his work with Tversky:

> Our research attempted to obtain a map of bounded rationality, by exploring the systematic biases that separate the beliefs that people have and the choices they make from the optimal beliefs and choices assumed in rational-agent models. The rational-agent model was our starting point and the main source of our null hypotheses, but Tversky and I viewed our research primarily as a contribution to psychology, with a possible contribution to economics as a secondary benefit. (2003b: 1449)

In other words, while working in what they viewed as a psychological research programme, they used a theory of rational choice as a benchmark for identifying mechanisms that might cause deviations from that theory.

Compare how Allais uses his paradoxes. He constructs a pair of artificially controlled decision tasks that are equivalent according the currently received theory of choice under uncertainty, EUT. He finds a significant difference in individuals' responses to those tasks that can be interpreted as exhibiting a mechanism of preference complementarity. His claim is not that people in fact behave according to EUT *except for* the effects of this mechanism; it is that EUT fails to take account of the mechanism. This is a one-deviation experimental methodology that follows the same logic as Kahneman and Tversky's investigation of intuitive judgements.

Perhaps Kahneman's criticism is not of one-deviation experimental research, but of one-deviation *theoretical models*. However, if models are interpreted as attempts to isolate particular mechanisms (or 'capacities') that operate alongside others in the world, models and experiments have many similarities (Cartwright, 2009; Mäki, 2009; Sitzia and Sugden, 2011). To move from the discovery of an experimental exhibit to a hypothesis that explains it, one

needs to specify a mechanism that, were it to operate, would produce not only the specific behavioural regularity of the exhibit but also other effects that can be the subject of empirical tests. That requires the construction of a model of the implications of the hypothesised effect when other factors are held constant—in other words, a one-deviation model. The methodological strategy we have just described was in fact the methodology of the programme of research into non-expected utility that responded to the Allais paradoxes. Theorists proposed alternative mechanisms, each of which explained those paradoxes but in different ways. This produced a family of alternative one-deviation models whose implications were then tested.[15]

Understood in this way, one-deviation models can be seen as part of a cumulative programme of scientific discovery. Progress is made by adding to the stock of evidence of mechanisms that affect human behaviour. It is true that the resulting collection of knowledge will not in itself take the form of a unified theory of human behaviour. But Kahneman says exactly the same about the programme of investigating intuitive judgements. Recall that, for him, a unified theory of a scientific domain is a thing of beauty and an object of aspiration, but as long as you are making discoveries you are making progress (see Section 2 above).

Research programmes using one-deviation modelling strategies can sometimes create progress by unexpectedly discovering evidence of mechanisms different from those being studied. Our 'That was Amos' story (Section 5 above) provides an illustration. The experiment we described to Kahneman grew out of our work with Loomes, developing and testing a theory that proposes an apparently reasonable motive that a decision maker might have—a concern for how what arises from a choice compares with what might otherwise have been. Loomes and Sugden (1982) showed theoretically that extending the utility-maximising framework of EUT by allowing this regret motive to operate provided a possible explanation for phenomena that multiple other theories were seeking to explain. This one-deviation model also generated novel predictions about systematic violations of the EUT properties of transitivity and stochastic dominance. Subsequent experimental testing produced evidence of these violations, but also showed that much of that evidence was due to the previously unobserved phenomenon of *event splitting*—a tendency for decision makers to assign extra weight to an event when it is decomposed into multiple sub-events (Starmer and Sugden, 1992). This phenomenon is a key driver of the intransitivity found in Starmer's (1999) experiment that we teased Kahneman about. Event splitting is naturally interpreted as arising from intuitive judgements about probability that do not conform to the conventional rules of probability. The fact that such findings can be generated by using the one-deviation methodology of theory and experiment suggests that the relationship between behavioural economics and psychology is more of a two-way street than Kahneman thought.

8. Conclusion

The objective of our paper was to not to evaluate Kahneman's contribution to economics but to investigate what he himself thought about it. However, we do not want to end without

[15] See Camerer (1992) and Starmer (2000) for reviews of the early work in this programme.

recording our conviction that Kahneman's contribution to economics has been enormously valuable. He and Tversky have done more than anyone else to make economists aware of the role of intuitive judgements in economic decision making, and particularly of the importance—we are tempted to say, the ubiquity—of reference-dependence in human judgement. In addition to that, they have done more than anyone else to persuade economists to use experimental methods to develop and test decision theories. The Nobel citation honoured Kahneman and Smith as the respective pioneers of 'psychological economics' and 'experimental economics', but as economists who have been experimentalists since the 1980s, we would prefer to say that Kahneman's and Smith's achievements include pioneering two different branches of experimental economics.

In relation to the objective of our paper, our assessment is that the gap between Kahneman's programme of psychological research and the research programme of behavioural economics—in which we include the work of those economists who, from the 1950s, investigated and tried to explain patterns of behaviour that contravened received theories of rational choice—is smaller than he thought. In our view, Kahneman was too willing to accept Tversky's mental picture of economics as committed to a normative model of the ideally rational agent. It is an ironic fact that theoretically-inclined commentators on economics—most commonly from mathematics or analytical philosophy, but in Tversky's case, from mathematical psychology—find the purest forms of economic theory more engaging than the work that the vast majority of economists actually do. Such commentators often criticise economics as a whole for its supposed commitment to abstract ideas developed by the discipline's pure theorists. The truth is that most economics investigates empirical questions, using theories that are gradually—some might say, too gradually—adapted in the light of empirical discoveries.

This is not to deny that, as Kahneman always maintained, psychology and economics have different overarching research objectives and that their methodologies differ in some significant respects. But there are important areas of enquiry in which economics and psychology pursue cross-cutting objectives while recognising the same observational data and employing similar methodologies, including the same standards for judging the success of empirical explanations. The Allais paradoxes, the Ellsberg paradox, the preference reversal phenomenon, the endowment effect and the anomaly in probabilistic reasoning exhibited in the 'Linda' task are problems on the agendas of both disciplines. In these areas of overlap, the customary methodologies of psychology and economics are complementary. In the research programme of behavioural economics to which Kahneman has contributed so much, each discipline can learn, and has learned, from the other.

References

Allais, Maurice (1953). Le comportement de l'homme rationnel devant le risque; critique des postulats et axiomes de l'ecole Americaine. *Econometrica* 21: 503–546.

Bardsley, Nicholas, Robin Cubitt, Graham Loomes, Peter Moffatt, Chris Starmer and Robert Sugden (2010). *Experimental Economics: Rethinking the Rules*. Princeton University Press.

Battigalli, Pierpaolo and Martin Dufwenberg (2007). Guilt in games. *American Economic Review: Papers and Proceedings* 97: 171–76.

Bell, David (1982). Regret in decision making under uncertainty. *Operations Research* 30: 961–981.

Bruni, Luigino and Robert Sugden (2007). The road not taken: how psychology was removed from economics, and how it might be brought back. *Economic Journal* 117: 146–173.

Camerer, Colin, F. (1992). Recent tests of generalizations of expected utility theory. In Ward Edwards (ed.), *Utility: Theories Measurement and Applications,* pp. 207-51, Kluwer.

Cartwright, Nancy (2009). If no capacities then no credible worlds. But can models reveal capacities? *Erkenntnis* 70(1): 45–58.

Chew, Soo Hong and Kenneth MacCrimmon (1979). Alpha-nu choice theory: a generalization of expected utility theory. Working Paper 669, University of British Columbia.

Ellsberg, Daniel (1961). Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics* 75(4): 643–669.

Fehr, Ernst and Klaus Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114(3): 817–868.

Friedman, Milton and Leonard Savage (1948). The utility analysis of choices involving risk. *Journal of Political Economy* 56(August): 279-304

Hagen, Ole (1979). Towards a positive theory of preferences under risk. In Maurice Allais and Ole Hagen (eds), *Expected Utility Hypotheses and the Allais Paradox*. Dordrecht: Reidel.

Handa, Jagdish (1977). Risk, probabilities, and a new theory of cardinal utility. *Journal of Political Economy* 85 (1): 97–122

Harless, David and Colin Camerer (1994). The predictive utility of generalized expected utility theories. *Econometrica* 62(6): 1251–1289.

Herfeld, Catherine (forthcoming). Psychological perspectives on rational choice theory: Daniel Kahneman. In Catherine Herfeld (ed.), *Conversations on Rational Choice*, Section 4.2. Cambridge University Press, forthcoming.

Hey, John and Chris Orme (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62(6): 1291–1326.

Kahneman, Daniel and Amos Tversky (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47(2): 263–291.

Kahneman, Daniel (2002). Maps of bounded rationality: a perspective on intuitive judgment and choice. Prize Lecture delivered on 8 December 2002. https://www.nobelprize.org/uploads/2018/06/kahnemann-lecture.pdf

Kahneman, Daniel (2003a). A psychological perspective on economics. *American Economic Review: Papers and Proceedings* 93(2): 163–168.

Kahneman, Daniel (2003b). Maps of bounded rationality: psychology for behavioral economics. *American Economic Review* 93(5): 1449–1475.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kim, Han, Adair Morse and Luigi Zingales (2006). What has mattered to economics since 1970. *Journal of Economic Perspectives* 20(4): 189–202.

Laibson, David (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112(2): 443–477.

Lichtenstein, Sarah and Paul Slovic (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89: 46–55.

Loomes, Graham and Robert Sugden (1982). Regret theory: an alternative theory of rational choice under uncertainty. *Economic Journal* 92(368): 805–824.

Loomes, Graham and Robert Sugden (1986). Disappointment and dynamic consistency in choice under uncertainty. *Review of Economic Studies* 53(2): 271–282.

Machina, Mark (1982). 'Expected utility' analysis without the independence axiom. *Econometrica* 50: 277-323.

Mäki, Uskali (2009). MISSing the world: models as isolations and credible surrogate systems. *Erkenntnis* 70(1): 29–43.

Markowitz, Harry (1952). The utility of wealth. *Journal of Political Economy* 60(2): 151–158.

Mosteller, Frederick and Philip Nogee (1951). An experimental measurement of utility. *Journal of Political Economy* 59(5): 371–404.

Munro, Alistair and Robert Sugden (2003). On the theory of reference-dependent preferences. *Journal of Economic Behavior and Organization* 50: 407–28.

NobelPrize.org (2002). Press release. https://www.nobelprize.org/prizes/economic-sciences/2002/press-release/

Preston, Malcolm and Philip Baratta (1948). An experimental study of the auction-value of an uncertain outcome. *American Journal of Psychology* 61(2): 183–193. .

Quiggin, John (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, *3*(4), 323–343.

Rabin, Matthew (1993). Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281–1302.

Savage, Leonard (1954). *The Foundations of Statistics*. Wiley. (Page references to 1972 edition, published by Dover [Mineola, New York].)

Sent, Esther-Mirjam (2004). Behavioral economics: how psychology made its (limited) way back into economics. *History of Political Economy* 36 (4): 735–760.

Simon, Herbert (1955). A behavioral model of rational choice. *Quarterly Journal of Economics* 69(1): 99–118.

Sitzia, Stefania and Robert Sugden (2011). Implementing theoretical models in the laboratory, and what this can and cannot achieve. *Journal of Economic Methodology* 18(4): 323–343.

Slovic, Paul, Melissa Finucane, Ellen Peters and Donald MacGregor (2007). The affect heuristic.

Starmer, Chris (1999). Cycling with rules of thumb: an experimental test for a new form of non-transitive behaviour. *Theory and Decision* 46(2): 141–158.

Starmer, Chris (2000). Developments in non-expected utility: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38(June): 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(1): 79–102.

Starmer, Chris and Robert Sugden (1993). Testing for juxtaposition and event-splitting effects. *Journal of Risk* and *Uncertainty* 6(3): 235–254.

Sugden, Robert (1993). An axiomatic foundation for regret theory. *Journal of Economic Theory* 60(1): 159–80.

Sugden, Robert (2005). Experiments as models and experiments as tests. *Journal of Economic Methodology* 12: 291–302.

Sugden, Robert (2024). Daniel Kahneman and the concept of the true self. Forthcoming in *Behavioural Public Policy*. Online access: https://doi.org/10.1017/bpp.2024.39

Thaler, Richard (2015). *Misbehaving: How Economics Became Behavioural*. London: Allen Lane.

Tversky, Amos and Daniel Kahneman (1986). Rational choice and the framing of decisions. *Journal of Business* 59(4), Part 2: S251–78.

Tversky, Amos and Daniel Kahneman (1991). Loss aversion in riskless choice: a reference-dependent model. *Quarterly Journal of Economics* 106: 1039–1061.

Tversky, Amos and Daniel Kahneman (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5: 297–323.