



Savage's response to Allais as Broomean reasoning

Franz Dietrich , Antonios Staras & Robert Sugden

To cite this article: Franz Dietrich , Antonios Staras & Robert Sugden (2020): Savage's response to Allais as Broomean reasoning, Journal of Economic Methodology, DOI: [10.1080/1350178X.2020.1857424](https://doi.org/10.1080/1350178X.2020.1857424)

To link to this article: <https://doi.org/10.1080/1350178X.2020.1857424>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 24 Dec 2020.



Submit your article to this journal [↗](#)



Article views: 148



View related articles [↗](#)



View Crossmark data [↗](#)

Savage's response to Allais as Broomean reasoning

Franz Dietrich^a, Antonios Staras^b and Robert Sugden^c

^aParis School of Economics and CNRS (Centre National de la Recherche Scientifique), Paris, France; ^bInstitute of Economics, Cardiff University, Cardiff, UK; ^cSchool of Economics, University of East Anglia, Norwich, UK

ABSTRACT

Savage famously contravened his own theory when first confronting the Allais Paradox, but then convinced himself that he had made an error. We examine the formal structure of Savage's 'error-correcting' reasoning in the light of (i) behavioural economists' claims to identify the latent preferences of individuals who violate conventional rationality requirements and (ii) Broome's critique of arguments which presuppose that rationality requirements can be achieved through reasoning. We argue that Savage's reasoning is not vulnerable to Broome's critique, but does not provide support for the view that behavioural scientists can identify and counteract errors in people's choices or preferences.

KEYWORDS

Savage; Allais Paradox; Broome; rationality; reasoning; behavioural economics



One of the most famous incidents in the history of behavioural economics took place in Paris in May 1952.¹ The Centre National de la Recherche Scientifique (CNRS) was holding a symposium on 'Foundations and applications of the theory of risk-bearing'. At the meeting, Leonard Savage presented the axiomatisation of subjective expected utility theory that would form the core of *The Foundations of Statistics* (Savage, 1954) and become the canonical statement of the theory of rational individual choice. Maurice Allais presented an initial version of a critique of expected utility theory that would subsequently be published in *Econometrica* and become a founding text of behavioural economics (Allais, 1953). During a lunch break, Allais invited Savage to respond to two hypothetical decision 'situations' requiring choices between pairs of gambles. Taken together, Savage's responses were inconsistent with 'P2', one of the fundamental axioms of his own theory. In *The Foundations of Statistics*, Savage refers to this episode and explains the inferences that he has drawn from it. He writes:

When the two situations were first presented, I immediately expressed preference for Gamble 1 as opposed to Gamble 2 and for Gamble 4 as opposed to Gamble 3 [thereby contravening P2], and I still feel an intuitive attraction to those preferences. But I have since accepted the following way of looking at the two situations ...

He explains how, after looking at the situations in a particular way, he came to prefer Gamble 3 to Gamble 4, and concludes:

It seems to me that in reversing my preference between Gambles 3 and 4 I have corrected an error. (1954, p. 103)

Our paper examines the formal structure of the mental process by which, on Savage's account, he comes to satisfy P2 and, in doing so, to correct an error. Despite the amount of attention that the interchange between Allais and Savage has already received, we believe that new insights can be gained by looking at Savage's error-correcting reasoning in the light of recent developments in behavioural economics and cognitive science and in the philosophy of rationality and reasoning.

CONTACT Robert Sugden  r.sugden@uea.ac.uk  School of Economics, University of East Anglia, Norwich NR4 7TJ, UK

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

The idea that many frequently-observed contraventions of rational choice theory are the result of errors that the relevant decision makers would wish to correct, and would be capable of correcting, is a common theme in current discussions of the findings of behavioural economics. It is used in two significant ways. In the literature of *normative decision theory*, it is used to commend (or ‘preach’) the axioms of rational choice theory as attainable standards of rationality while recognising that actual choice behaviour often contravenes them. In the literature of *behavioural welfare economics*, it is used to justify apparently paternalistic public policies (such as ‘nudging’) on the grounds that those policies help people to avoid erroneous choices. Although these two literatures are concerned with different types of *response to errors* (in one case, correction by the individual herself, in the other, public policies that aim to change behaviour), they both seem to presuppose the existence of some mode of error-free reasoning by which individuals can construct or amend their preferences so as to satisfy the axioms of rational choice theory. Since standard versions of rational choice theory lack any model of reasoning processes, the truth of this presupposition is difficult to assess.

In a major philosophical work that has received insufficient attention from economists and decision theorists, Broome (2013) examines the relationship between rationality and reasoning. Broome points out that many writers on rationality ‘seem to think that they have finished their job when they have described requirements of rationality’. He suggests that these writers ‘must believe that, starting from knowledge of a particular requirement, you can reason your way actively to satisfying that requirement’ (2013, pp. 208–209), and asks whether that belief is justified. In addressing that question, Broome develops a model of a person’s mental states that can represent both requirements of rationality and processes of reasoning.² By identifying serious problems in modes of (purported) reasoning that proceed from beliefs about what rationality requires to the satisfaction of rationality requirements, he casts doubt on the effectiveness of attempts to preach normative decision theory. Broome’s work suggests the need for further investigation of how decision theorists and behavioural economists use the concept of error, and of how, according to these accounts, individuals can correct their own errors; and it provides a conceptual framework for such an investigation.

Seen against the background of these intellectual developments, Savage’s response to Allais is of great interest. The formal contribution of *The Foundations of Statistics* to rational choice theory is an analysis of coherence principles that apply to a person’s preferences between ‘acts’ under uncertainty. Almost all the supporting argument is structured to justify those principles *as requirements of rationality*; there is no formal analysis of reasoning processes. To this extent, Savage falls into Broome’s class of writers on rationality who seem to think it sufficient to describe requirements. But in reporting his considered response to Allais’s challenge, Savage gives an informal account of how, by reasoning outside his axiomatic model, he came to satisfy one of its requirements and, in doing so, to correct an error. Given that Savage is one of the all-time greats of decision theory, this account is a natural place to look for clues about how violations of rationality requirements might be understood as errors that can be corrected by reasoning.

Using the formalisation of Broome’s account of rationality and reasoning presented by Dietrich et al. (2019), and drawing on the psychological distinction between the automatic mental processes of ‘System 1’ and the conscious reasoning processes of ‘System 2’ (Kahneman, 2003; Wason & Evans, 1974), we reconstruct Savage’s informal reasoning as an explicit sequence of mental operations, in order to address two main issues.

The first issue relates to normative decision theory, and concerns the internal logic of Savage’s reasoning. Does this reasoning avoid Broome’s objections? We will argue that the answer is ‘Yes’: it provides an example of how a person might reason his way towards satisfying a rationality requirement. It thereby provides an example of how rationality requirements can form attainable normative standards, and of how normative decision theory can be preached successfully – in Savage’s case, *to oneself*, but more generally *to others*. Significantly, however, Savage’s reasoning – unlike that invoked by the writers whom Broome criticises – does not proceed from premises about what rationality requires. In this respect, our analysis supports Broome’s account of the relationship between rationality and reasoning.

The second issue relates to behavioural welfare economics. Does Savage's mode of reasoning confirm the legitimacy of the concept of correctable error that behavioural economists use when making policy recommendations? We will argue that the answer is 'No'. Savage's concept of error correction is coherent, but it is not the concept that behavioural welfare economics invokes. This negative conclusion adds weight to doubts about the methodological foundations of behavioural welfare economics that have been expressed in previous literature (e.g. Infante et al., 2016; Rizzo & Whitman, 2020, pp. 46–52). Up to now, those critiques have had little impact on the practice of behavioural economics, perhaps because the idea that violations of rationality requirements are errors is very intuitive. By distinguishing between two different concepts of error, our Broomean analysis shows that that intuition does not necessarily support the methods of behavioural welfare economics.

The structure of our paper is as follows. In Section 1, we discuss how the concept of error has been used in normative decision theory and behavioural welfare economics. In Section 2, we summarise Broome's account of rationality and reasoning, to the extent relevant here. In Section 3, we examine how Savage uses his informally stated Sure-thing Principle in justifying P2 as a requirement of rationality, and look at Savage's account of how he violated P2 when he first faced Allais's situations. Section 4 presents our reconstruction of the reasoning by which Savage revised his preferences over Allais's gambles, and shows that this reasoning is not vulnerable to Broome's objections. In Section 5, we explain the fundamental differences between Savage's concept of error and the concept used in behavioural economics. Section 6 concludes.

1. The concept of error in normative decision theory and behavioural welfare economics

The idea that violations of coherence principles of rational choice theory are errors appears in one of the earliest and most important contributions to behavioural economics – the paper in which Kahneman and Tversky (1979) propose *prospect theory*. Kahneman and Tversky begin by presenting a variety of experimental results which 'appear to invalidate expected utility theory as a descriptive model' (p. 274). These results include an instance of the common consequence effect – the effect revealed in Savage's initial response to Allais's situations. Kahneman and Tversky then propose prospect theory as an explanation of their experimental results. In the final paragraph of this 'Theory' section, they summarise the differences between prospect theory and expected utility theory, and conclude:

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)

Taken at face value, the ideas expressed in this passage are remarkably similar to those in Savage's response to the common consequence effect. The phenomena that Kahneman and Tversky describe as 'normatively unacceptable' are violations of rationality requirements of expected utility theory. According to Kahneman and Tversky, a reasonable decision maker would wish to satisfy those requirements. The decision maker's failure to do so in certain situations is a predictable consequence of human psychology, but it is also an error that he would want to correct, and *could* correct if (like Savage) he became aware that he was violating a rationality requirement. Unlike Savage, however, Kahneman and Tversky do not describe any process of reasoning that a decision maker might use to correct such errors.

1.1. The concept of error in the preaching of normative decision theory

In the passage we quoted above, Kahneman and Tversky are expressing a view of the normative status of expected utility theory that is common among decision theorists, and which has been

stated particularly clearly by Gilboa (2010, pp. 3–4). Expressing apparent regret that ‘phenomenally elegant classical decision theory’ has proved to be so easily disconfirmed in simple experiments like Kahneman and Tversky’s, Gilboa asks what we (presumably, decision theorists) should do. The approach he favours is not to revise the theory, but ‘to go out and preach our classical theories, that is, to use them as normative ones’, thus trying to ‘bring reality closer to the theory’. According to Gilboa’s definition, ‘a mode of behaviour is rational for a given decision maker if, when confronted with the analysis of her behaviour, the decision maker does not wish to change it’ (2010, p. 3). We take it that by ‘the analysis’, Gilboa means a theoretical analysis that uses the principles of classical decision theory, and that in talking about ‘wishing to change’ a choice, he is imagining a person who has made a provisional choice but revises it after those principles have been preached to her. Then:

If decision makers become convinced once the theory has been explained to them and they then wish to change their choices (that is, if their choices were irrational to them), we may declare the classical theory successful as a normative one. It would indeed be reasonable to preach the classical theory and help decision makers make better (as judged by themselves) decisions. (2010, p. 4)

Notice the parallel with Savage: after explaining his theory to himself, Savage wished to change a choice that he had made in one of Allais’s situations. For Gilboa, the test of the normative status of classical decision theory is whether it can be preached successfully. But if in fact the theory *can* be preached successfully, Gilboa treats violations of it as errors that those decision makers would want to correct and are capable of correcting. Implicitly, he is assuming that knowledge and understanding of rationality requirements can trigger some process of error-correcting reasoning. But, unlike Savage, he offers no description of such a process.

1.2. The concept of error in behavioural welfare economics

From the first years of behavioural economics, the question of whether ‘anomalies’ (i.e. patterns of choice that violate standard rationality requirements) should be interpreted as errors has remained open.³ However, questions about error have become more significant since the early 2000s, when behavioural economists began to consider the normative implications of their findings, and to propose public policies based on ‘behavioural insights’. Much recent work in normative behavioural economics is premised on the assumption that (as Thaler and Sunstein put it in their book *Nudge*) ‘individuals make pretty bad decisions – decisions that they would not have made if they had paid full attention and possessed complete information, unlimited cognitive abilities, and complete self-control’. Much of the evidence that is presented in support of this assumption takes the form of violations of conventional rationality requirements. Public policy, it is argued, should be designed to counter individuals’ tendency to make bad decisions. Crucially, its aim should be to ‘make choosers better off, *as judged by themselves*’ (Thaler & Sunstein, 2008, p. 5; italics in original). Or, as Thaler (2015, p. 326) puts it, he and Sunstein ‘just want to reduce what people would themselves call errors’. The implication is that an individual’s judgements about what makes her better off are expressed in the choices she herself would make in the absence of, or after correcting for, errors induced by inadequate attention, information, cognitive ability and self-control. By abstracting from the effects of these errors on observable choice behaviour, analysts can reconstruct individuals’ *latent* (or ‘underlying’ or ‘true’) preferences. The satisfaction of (assumedly consistent) latent preferences can then be used as a criterion for guiding public policy.

The fundamental hypothesis that violations of rationality requirements result from interactions between error-inducing psychological processes and coherent latent preferences and beliefs appears in many other contributions to normative behavioural economics (e.g. Bershears et al., 2008; Bleichrodt et al., 2001; Camerer et al., 2003; Köszegi & Rabin, 2008; Manzini & Mariotti, 2014; Salant & Rubinstein, 2008). Following Infante et al. (2016), we will use the term *behavioural welfare economics* for all forms of normative economics with this feature. It is perhaps not surprising that so many economists have favoured this approach when responding to empirical findings that

show that individuals' revealed preferences are often highly sensitive to features of decision situations that seem to have no relevance for welfare. By reconceptualising preference-satisfaction in terms of latent rather than revealed preferences, economists can continue to use much of the formal apparatus of traditional welfare economics. And by defining a person's latent preferences as the preferences she would reveal in the absence of error, they can continue to uphold the principle of respecting each person's judgements about her own good – the idea that is expressed in the 'as judged by themselves' clause that is now widely used by behavioural economists. In taking this approach, behavioural economists seem to implicitly assume the existence of some form of error-free reasoning that can be subverted by psychological mechanisms. Advocates of behavioural welfare economics have generally failed to explain the psychological basis of this reasoning, or what form it might take.⁴

Some advocates of behavioural welfare economics represent the relationship between error and latent preference in terms of the 'System 1/ System 2' classification of mental processes proposed by Wason and Evans (1974) and developed by Kahneman (2003). System 1 is fast and automatic, and generates impressions, intuitions, feelings and impulses. System 2 is slow and under conscious control; using it is effortful. Operating on the outputs of System 1, System 2 constructs explicit thoughts in an orderly way. The distinction between the two systems is used by Thaler and Sunstein (2008, pp. 19–39) and Kahneman (2011) as a way of organising behavioural findings. In both cases, the suggestion is that System 1 is liable to induce unconsidered preferences and judgements that are systematically biased, and that System 2 is capable of correcting. As Kahneman and Sunstein (2006, p. 92) put it: 'System 1 quickly proposes intuitive answers to judgment problems as they arise, and System 2 monitors the quality of these proposals, which it may endorse, correct, or override'. However, little is said about the form that System 2 reasoning might take, or about any limitations it might have.

It is crucial to the methodology of behavioural welfare economics that, in the process of retrieving an individual's latent preferences, the 'errors' that are screened out are errors *according to that individual's own standards* – in Thaler's expression, 'what people would themselves call errors'. If policy interventions are to be directed at people in general (rather than tailored to the needs of specific individuals), the concept of error must be intersubjective – error according to standards that almost everyone accepts. It is also crucial to distinguish between this concept of error and the violation of rationality conditions. For example, imagine a person who chooses x from $\{x, y\}$, y from $\{y, z\}$ and z from $\{x, z\}$. If we are to attribute transitive latent preferences to this person, and if we assume that choices reveal strict preferences, at least one of the three revealed preferences must be 'corrected'. But a behavioural welfare economist needs to know *which* preference to correct. That requires the identification of an error of mental processing in the formation of a *specific* preference. Identifying the violation of a rationality requirement is not enough.

2. Achieving rationality through reasoning: Broome's scepticism

Broome (2013) asks whether rationality requirements, such as those encapsulated in the axioms of decision theory, can be achieved by reasoning. He expresses surprise that many writers on rationality seem to assume that, if a person comes to believe that some specific principle is a requirement of rationality, she is thereby able to reason her way to satisfying that requirement. In questioning this assumption, he develops a theory of a person's rationality and reasoning. Both concepts are located in mental states, but play fundamentally different roles. We begin by presenting those parts of the theory needed for the purposes of this paper, using where necessary the formalisation developed by Dietrich et al. (2019). We then explain why, for Broome, achieving rationality through reasoning is not straightforward.

At any given time, an individual (in Broome's language, 'you') has a set of *mental states*. A mental state is an attitude that you hold towards something (the *content* of that attitude). The simplest kind of mental state is a pair (p, a) where p , the content, is a proposition and a is an *attitude type* that it is

possible to hold towards p . For example, (q, belief) is the mental state of believing that q is the case; $(I \text{ do } x, \text{intention})$ is the mental state of intending to do x . There are also mental states whose attitude types are held towards ordered pairs of propositions. For example, $(I \text{ get } x, I \text{ get } y, \text{preference})$ is the mental state of strictly preferring getting x to getting y ; $(I \text{ get } x, I \text{ get } y, \text{indifference})$ is the mental state of being indifferent between getting x and getting y . Let M be the set of all possible mental states. At any given time, you have a subset C of M , called a (*mental*) *constitution*.

A *requirement* is a condition on your constitution, and hence restricts the mental states that you can hold at the same time. Formally, it identifies a 'permissible' subset of the set 2^M of all possible constitutions. Constitutions in this subset *satisfy* the requirement; the others *violate* it. Our concern is with requirements that are interpreted as requirements of *rationality*. An example is the principle that strict preferences are transitive. It can be written as a schema of rationality requirements: for any triple of options x, y, z , there is the requirement that, if your constitution contains $(I \text{ get } x, I \text{ get } y, \text{preference})$ and $(I \text{ get } y, I \text{ get } z, \text{preference})$, then it also contains $(I \text{ get } x, I \text{ get } z, \text{preference})$.⁵

Rationality, as understood by Broome (2013, p. 152), 'requires your mental states to be properly related to one another'. It has nothing to say about how mental states come into or go out of existence. In contrast, *reasoning* is a causal psychological process that creates new mental states. It is an activity under your conscious control. Thus, in terms of the System 1/ System 2 distinction, reasoning is a System 2 operation. However, reasoning is not the only mental process that creates mental states; mental states can be created (and also eliminated) by the automatic psychological processes of System 1.

Broome (2013, p. 234) describes reasoning as 'a rule-governed operation on the contents of your conscious attitudes'. Your mental experience of reasoning consists in calling to mind a set of premise-attitudes and then finding (in a way that can be expressed as 'So ...') that some conclusion-attitude follows from them. In 'calling an attitude to mind', you bring its content into your conscious mind in a way that allows you to use it in reasoning (2013, pp. 222–224). The implication is that, at any given time, your conscious attitudes make up a subset of the (probably much larger) set of mental states that you have already formed and are *capable of* calling to mind. The latter set is your constitution. The sense of 'following from' is that of (implicitly or explicitly) being guided by some rule, in a way that 'seems right to you' (2013, pp. 237–8). For a mental process to be reasoning, it is not necessary that you are consciously aware of the rule you are following, or even of the concept of a rule (2013, p. 233). It will be convenient to apply the concept of *seeming right* to rules whose individual applications seem right to you; but, strictly speaking, what seems right to you is the mental process and not the rule itself.

Formally, a *reasoning rule* is a pair (P, m) where P is a set of *premise* mental states and m is a *conclusion* mental state. (Analogously with our treatment of requirements, we sometimes use the term 'rule' as a shorthand for 'schema of rules', by treating P and m as variables or parameters of a schema.⁶) For example, suppose the rule that allows you to create the conclusion mental state $(I \text{ get } x, I \text{ get } z, \text{preference})$ from the pair of premise mental states $(I \text{ get } x, I \text{ get } y, \text{preference})$ and $(I \text{ get } y, I \text{ get } z, \text{indifference})$ seems right to you, and that you already have these premise mental states. Then you can reason in a way that you can put into words as: 'Rather x than y . Just as well y as z . So, rather x than z '. This causes you to have the mental state $(I \text{ get } x, I \text{ get } z, \text{preference})$. In your internal language, 'Rather x than y ' expresses the *marked content* of your preference for x over y , i.e. the content $(I \text{ get } x, I \text{ get } y)$ marked by the attitude type of preference. Similarly, 'Just as well y as z ' expresses the marked content of $(I \text{ get } y, I \text{ get } z, \text{indifference})$. Beliefs are special mental states in that they need no linguistic marker: our internal language expresses the marked content of $(It \text{ rains}, \text{belief})$ simply as 'It rains'. Notice that your reasoning is about the *contents* of your attitudes, not about the attitudes themselves. Thus, in the example, the conclusion of your reasoning is 'Rather x than z ' and not 'I prefer x to z '. The latter would express the marked content of the *second order* mental state $(I \text{ prefer } x \text{ to } z, \text{belief})$. (A mental state of yours is second-order if its content refers to one or more of your own mental states.)

A person might make a mistake in reasoning. Recall that, both in the preaching of normative decision theory and in behavioural welfare economics, the relevant errors are ones that the decision maker would want to correct – i.e. errors according to the decision maker's own standards. Broome (2013, pp. 232–234, 237–240) uses a similar concept of error when referring to cases in which you have a 'steady disposition' for a certain rule to seem right to you, but you make a mistake in trying to follow it. For example, you might try to follow the rules of arithmetic in adding 11,866 to 14,977, but make an arithmetic error. The mental process that leads to the conclusion 'So, the sum is 26,853' seems right to you at the time, and therefore counts as reasoning, but it involves a mistake according to your own standards of rightness. We set aside a different type of reasoning error: rules can be 'correct' or 'incorrect' by a universal or intersubjective standard, following Broome (2013, p. 240). We consider only errors in applying your rules, not 'incorrectness' of your rules.

We now have the resources to explain some of the difficulties involved in trying to reason from a belief in a rationality requirement to the satisfaction of that requirement.

The most obvious difficulty is that accepting a rationality requirement does not imply accepting a reasoning rule by which you can achieve that requirement. A reasoning rule may indeed be a means by which a rationality requirement can be achieved, but for that rule to be *your* rule – and thereby to be a rule that you can use – it must seem right to you. Believing that a rule is a means to some desirable end does not necessarily make the rule seem right. More fundamentally, there is nothing in Broome's model that allows you to reason towards having any specific rule *r*. Since '*r* is a rule of mine' expresses a proposition, it cannot be a mental state in itself; it can only be the *content* of some mental state, such as belief or intention. Believing or intending that *r* is a rule of yours is not the same thing as the rule seeming right to you.

One way of overcoming this difficulty might be to assume that you have some general second-order *enkratic* rule (strictly, a schema of such rules) that allows you to reason directly from the belief in the requirement to have a certain mental state to the intention to have it. For example, you might have the rule that allows you to create (*I have mental state m, intention*) from (*rationality requires that I have mental state m, belief*), where *m* is any mental state. But this gives rise to two new problems.

First, the requirements of decision theory typically do not declare individual mental states as unconditionally required; instead they have an *If ... , then ...* structure. In some cases, either the antecedent or the consequent (or both) is a proposition about the absence of mental states. For example, consider the requirement of preference completeness, which prescribes the presence of (*I get y, I get x, preference*) given the absence of (*I get x, I get y, preference*) and (*I get x, I get y, indifference*). Or consider the requirement of preference acyclicity, which prescribes the absence of (*I get z, I get x, preference*) given the presence of (*I get x, I get y, preference*) and (*I get y, I get z, preference*). But reasoning, in Broome's theory, cannot start from or conclude in the absence of mental states. As a result, as shown formally by Dietrich et al. (2019), reasoning is able to achieve *closedness requirements* (requirements prescribing presences given presences), but can neither achieve *consistency requirements* (requirements prescribing absences given presences) nor, to some extent, achieve *completeness requirements* (requirements prescribing presences given absences). Broome (2013, pp. 279–280) recognises that, although consistency requirements are essential to theories of rationality, they cannot be achieved by explicit reasoning. He argues that they are normally achieved by automatic psychological processes, but says little about what those processes might be. We will discuss this issue in Sections 4 and 5.

Second, having a certain mental state *m* arguably can be the content of an intention only if there is some psychological mechanism by which you can bring about *m* by a conscious mental act. Broome (2013) doubts whether there is any such direct mechanism for preferences if preferences are interpreted as beliefs about betterness (the interpretation he favours) or as comparative desires. However, he also considers (but dismisses as 'artificial') a concept of 'broad preference' that is arguably implicit in some presentations of decision theory. To have a broad preference for *x* over *y* is to be in a mental state that would typically cause you to choose *x* if the only available

options were x and y . Broome suggests that because some kinds of broad preference *are* intentions, it may be possible to reason directly from a belief that some broad preference is rationally required towards having that preference.

It is significant that the foregoing arguments are about *directly* bringing yourself to have particular mental states. Broome (2013, p. 213) recognises that in some circumstances you may be able to deliberately activate psychological mechanisms that will then cause you to form certain attitudes, but notes that such indirect processes are not reasoning. His examples of indirect processes are (in his words) ‘exotic’ – ‘enlisting the help of a hypnotist’, ‘undertaking a programme of self-persuasion’, and ‘in science fiction, [taking] a pill’. But there are everyday examples too. If you choose to browse in an up-market department store rather than a discount supermarket, you will expose yourself to psychological cues that activate desires for luxury goods. We will say more about such indirect processes later.

3. The Sure-thing Principle, Savage’s P2 axiom and Allais’s challenge

We now return to Savage and Allais, beginning by considering how Savage justifies P2 as a requirement of rationality.

Savage’s (1954) axiomatization of subjective expected utility theory (SEUT) is one of the greatest achievements of decision theory. His contribution is to show that the whole structure of SEUT is implied by a small set of axioms, formulated in a remarkably frugal theoretical framework and using only concepts that can be observed in choice behaviour. As an explanation of the value of an axiomatic approach, he draws an analogy with logic and says: ‘[T]he main use I would make of [my axioms] is normative, to police my own decisions for consistency and, where possible, to make complicated decisions depend on simpler ones’ (1954, p. 20).

Savage works in a theoretical framework whose primitives are *states* (of the world), *consequences*, *acts* (i.e. assignments of consequences to states), and (strict) *preferences*. Sets of states are *events*. A person’s preferences are described by a binary relation \leq among acts. If f and g are acts, $g \leq f$ is interpreted as ‘ g is not preferred to f ’ or, equivalently, as ‘ f is preferred or indifferent to g ’. The two interpretations are equivalent because Savage excludes the possibility of non-comparability.⁷ Savage’s P1 axiom requires that \leq satisfies completeness (i.e. for all acts f and g , $f \leq g$ or $g \leq f$) and transitivity (i.e. for all acts f , g and h , $[f \leq g \wedge g \leq h] \Rightarrow f \leq h$). Given the assumption that P1 holds, Savage’s formalism is equivalent to the modern one, since $g \leq f$ is equivalent to $f \geq g$ for a standard (complete and transitive) relation \geq of weak preference. However, Savage’s use of strict preference as a primitive reflects his determination to build his theory on observable concepts: ‘I think it of great importance that preference, and indifference, between f and g be determined, at least in principle, by decisions between acts and not by response to introspective questions’.

Elucidating the proposition that a person prefers f to g , he says: ‘Loosely speaking, what this means is that if he were required to decide between f and g , no other acts being available, he would decide on f ’ (1954, p. 17). ‘Loosely speaking’ and ‘at least in principle’ hint at a tension between two of Savage’s aims. On the one hand, he wants to develop a theory of decision that is open to empirical falsification by observable behaviour – not because he expects that its predictions will always be confirmed, but as what he sees as good scientific practice (1954, p. 20). On the other, he wants to provide normative justifications for specific principles of rational choice. Those justifications are addressed to actual human beings. He is telling his readers that they are required by principles of rationality to act *like* the abstract agents in his theory. In explaining why this is so, he quite properly steps outside his parsimonious behaviourist interpretations and appeals to the reader’s mental experience.⁸ In doing so, he effectively interprets his theory – with formal constructs such as preferences, conditional preferences, probabilities, and utilities – *as a representation of* a richer world that includes mental states as well as behaviour.

Our paper is primarily concerned with Savage's second axiom, P2. Using the definition that two acts f and g agree in an event E if, for all states $s \in E$, $f(s) = g(s)$, and using Ω to denote the set of all states of the world, we can write this postulate as:

Savage's P2. If acts f, g, f', g' and event E are such that:

- (1) in $\Omega \setminus E$, f agrees with g , and f' agrees with g' ,
 - (2) in E , f agrees with f' , and g agrees with g' ,
 - (3) $g \leq f$;
- then $g' \leq f'$.

As a preliminary to stating this axiom, Savage presents the *Sure-thing Principle*:

If the person would not prefer f to g , either knowing that the event B obtained, or knowing that the event $\sim B$ [i.e. the complement of B] obtained, then he does not prefer f to g . Moreover (provided he does not regard B as virtually impossible) if he would definitely prefer g to f , knowing that B obtained, and if he would not prefer f to g , knowing that B did not obtain, then he definitely prefers g to f . (1954, pp. 21–22)

Savage's illustration is the case of a businessman who is deciding whether to buy a piece of property. The outcome of that decision may depend on whether the upcoming US presidential election is won by the Republican candidate or by the Democrat (there are no others). The businessman 'finds' or 'decides' (Savage uses both expressions) that he would buy if he knew the Republican would win, and also that he would buy if he knew the Democrat would win. 'On the basis of' the Sure-thing Principle, he decides that, despite not knowing who will win, he should buy. Savage says of this principle: 'except possibly for the assumption of simple ordering [i.e. P1], I know of no other extralogical principle governing decisions that finds such ready acceptance' (1954, p. 21). Of course, this claim would be disputed by many modern theorists (and it had already been challenged by Allais when Savage was writing). But our concern here is with the role of the Sure-thing Principle in Savage's reasoning, not with its normative force.

It is important to recognise that the Sure-thing Principle, as stated by Savage, is not the same thing as P2; rather, it serves as part of his *justification* of that axiom. One difference is that the Sure-thing Principle, unlike P2, is a principle of what decision theorists would now call 'event-wise dominance'. More importantly, and as Savage (1954, p. 22) notes explicitly, the Sure-thing Principle refers to two concepts that are not defined in terms of his primitives.

The first of these concepts is 'knowing that' some event obtains. To a careful reader of *The Foundations of Statistics*, it will become clear that 'not preferring f to g , knowing that B obtains' is the informal counterpart of the ternary relation $f \leq g$ given B that Savage (1954, p. 22) defines within his formal system. But that definition depends on the assumption that P2 holds,⁹ an assumption that Savage cannot make when using the Sure-thing Principle to explain or justify P2. The second such concept is that of an event regarded as 'virtually impossible'. This also has a formal counterpart definable in Savage's framework, namely a *null event* (which becomes a zero-probability event if preferences have an expected-utility representation). But once again the formal concept (here, null event) or its interpretation in terms of the informal concept (here, virtually impossible event) hinges on assuming P2. For Savage, the Sure-thing Principle is a pre-theoretic principle that is self-evident to ordinarily intelligent persons unfamiliar with decision theory. He describes it as 'a loose [principle] that suggests certain formal postulates well articulated with [i.e. connected to] P1' (1954, p. 22). P2 is one of the postulates suggested by this Principle.

In a discussion of a range of 'critical comments on utility', Savage (1954, pp. 101–103) addresses and rejects the possible criticism that 'Introspection about certain hypothetical decision situations suggests that the sure-thing principle, and, with it, the theory of utility are normatively unsatisfactory'. He asks his readers to consider an example of two decision situations. Initially, he describes these situations as they were presented to him by Allais, but with prizes in US dollars.¹⁰ This presentation is shown in Table 1.

Table 1. The decision situations as presented by Allais.

<i>Situation 1: Choose between</i>	
Gamble 1.	\$500,000 with probability 1; and
Gamble 2.	\$2,500,000 with probability 0.1, \$500,000 with probability 0.89, status quo with probability 0.01.
<i>Situation 2: Choose between</i>	
Gamble 3.	\$500,000 with probability 0.11, status quo with probability 0.89; and
Gamble 4.	\$2,500,000 with probability 0.1, status quo with probability 0.9.

Savage conjectures that many of his readers will prefer Gamble 1 to Gamble 2 and prefer Gamble 4 to Gamble 3, and that, for almost everyone, there will be some variant of the example that will induce preferences of this kind. These were the preferences that he expressed when first shown the two situations, and he still feels an intuitive attraction to them. Savage offers a psychologically plausible explanation for this phenomenon:

Many people prefer Gamble 1 to Gamble 2, because, speaking qualitatively, they do not find the chance of receiving a very large fortune in place of receiving a large fortune outright adequate compensation for even a small risk of being left in the status quo. Many of the same people prefer Gamble 4 to Gamble 3; because, speaking qualitatively, the chance of winning is nearly the same in both gambles, so the one with the much larger prize seems preferable.

Table 2 is one way of representing this understanding of the two situations. The language in which chances and outcomes are described is imprecise, but it is natural and intuitive.

The Allais Paradox preferences revealed by Savage are jointly inconsistent with SEUT. Savage demonstrates this fact algebraically, by treating the utilities of the three relevant consequences as unknown variables and showing that those preferences are not consistent with any combination of values of those variables. At this initial stage, he does not try to identify which of his axioms he has contravened, presumably because Allais presented the situations in a form that does not have a unique translation into the state/consequence structure in which Savage’s axioms are formulated. Nevertheless, since those axioms jointly imply SEUT, the algebraic analysis is sufficient to establish that he has contravened at least one of them.

Savage finds it unsettling that his preferences are inconsistent with a theory that he has defended as normative:

In general, a person who has tentatively accepted a normative theory must conscientiously study situations in which the theory seems to lead him astray; he must decide for each by reflection – deduction will typically be of little relevance – whether to retain his initial impression of the situation or to accept the implications of the theory for it. (1954, p. 102)

Notice that, faced with the question of whether he personally accepts a normative implication of his theory, Savage switches from theoretical reasoning to ‘reflection’. As long as he is merely policing his

Table 2. Savage’s initial mental representation of the decision situations.

<i>Situation 1: Choose between</i>			
Gamble 1		certain large prize	
Gamble 2	unlikely very large prize	likely large prize	very unlikely status quo
<i>Situation 2: Choose between</i>			
Gamble 3	unlikely large prize	likely status quo	
Gamble 4	unlikely very large prize	likely status quo	

decisions for consistency with a given set of axiomatic requirements, he has no need to engage in introspection: he needs only to observe his own decisions and to investigate their properties. But now he is asking whether he still accepts that his axioms are justified as requirements of rationality. He needs to switch to a mode of reasoning that can accommodate the pre-theoretic concepts that are invoked by the Sure-thing Principle. It is at this point in a reconstruction of Savage's reasoning that a Broomean theoretical framework becomes useful.

4. A Broomean representation of Savage's 'reflections'

Our objective in this section is to reconstruct the mental process by which Savage arrived at preferences over Allais's situations that were consistent with P2, and to assess this in the light of Broome's catalogue of the difficulties of achieving rationality through reasoning. We first give representations (within Dietrich et al.'s Broomean framework) of the mental states, rationality requirements and reasoning rules that feature in Savage's argument.

We can define states of the world, events, consequences and acts as in Savage's theory. It would be natural to represent Savage's introspective concepts of preference and indifference between acts as Broomean attitude types. However, our reconstruction of Savage's pre-theoretic reasoning requires some other attitude types. Specifically, we need to represent three concepts that are not among Savage's formal primitives – 'is not virtually impossible', 'preferring ... knowing that ...', and 'indifferent between ... knowing that ...' (see Section 3 above). It turns out that ordinary 'unconditional' preference and indifference can be represented as special cases of the second and third of these concepts, obtained for vacuous knowledge.¹¹

The first concept is simpler to deal with. We define an attitude type of *partial belief*, using that term as a shorthand for 'belief that is *at least* partial'. The domain of this attitude type (i.e. the set of propositions that can feature in the content of that attitude) is the set of propositions of the form '*E obtains*' (which we abbreviate to '*E*'), where *E* is an event. The marked content of (*E*, *partial belief*) is expressed by 'Maybe *E*'. This is our first-order Broomean representation of the attitude that Savage expresses as 'not regarding *E* as virtually impossible'.

The second concept is more difficult, because its intended meaning is not immediately obvious. Take the case of the businessman who would prefer buying (*f*) to not buying (*g*) if he knew that the Republican would win an election that has yet to take place. It is clear that this is a preference held at a time when the businessman does not know the election result: it is not a second-order belief about what he *will* prefer at a future date, if and when he knows that the Republican has won. It is not the material conditional 'If I know the Republican will win, then I prefer *f* to *g*', since the antecedent of that conditional is false. But nor is it the counterfactual conditional 'Were I to know that the Republican would win, I would prefer *f* to *g*'. Knowing the result of an election before it takes place would be a peculiar counterfactual. We suggest that Savage's intention is best represented in terms of a primitive first-order attitude that can be expressed as 'On the assumption that the Republican will win, I prefer *f* to *g*'.

We formalise this idea using two further attitude types. For each event $E \subseteq \Omega$, we define the two-place attitude types *E*-conditional preference, denoted \succ_E , and *E*-conditional indifference, denoted \sim_E . The domain of each of these attitude types is the set of propositions of the form '*I get f*' (which we abbreviate to '*f*') where *f* is an act. The mental state (f, g, \succ_E) is the attitude of strictly preferring *f* to *g*, on the assumption that *E* obtains. The marked content of this attitude is expressed by 'Assuming *E*, rather *f* than *g*'. The mental state (f, g, \sim_E) is the attitude of being indifferent between *f* and *g*, on the assumption that *E* obtains; its marked content is expressed by 'Assuming *E*, just as well *f* as *g*'. Ordinary preference and indifference between acts are identified with preference and indifference given the tautological event Ω , i.e. \succ_Ω and \sim_Ω respectively. The marked contents of (f, g, \succ_Ω) and (f, g, \sim_Ω) can be expressed simply as 'Rather *f* than *g*' and 'Just as well *f* as *g*'. For example, consider Savage's businessman. Let *E* be the event that the election is won by the Republican. According to Savage, the businessman has the initial mental states (f, g, \succ_E) and $(f, g, \succ_{\Omega \setminus E})$, and forms the new mental state (f, g, \succ_Ω) .

For our reconstruction of Savage's response to Allais, the only rationality requirements we need to consider are instances of the following schemas, whose variables are two acts and an event, respectively denoted f , g and E (despite the fixed meanings we sometimes give to these symbols):

Asymmetry of preference (AP): $(f, g, >_E) \in C \Rightarrow (g, f, >_E) \notin C$.

Sure-thing Principle (for combining preference and indifference) (STP): $[(f, g, >_E) \in C \wedge (f, g, \sim_{\Omega E}) \in C \wedge (E, \text{partial belief}) \in C] \Rightarrow (f, g, >_{\Omega}) \in C$.

We take these requirements to be ones that Savage believes to be self-evident to ordinarily intelligent persons.¹² It might be objected that AP is a conceptual necessity rather than a rationality requirement. Indeed, Savage (1954, p. 17) argues that *in his formal theory* 'the very meaning' of preference in terms of pairwise choice implies that a person cannot simultaneously prefer f to g and g to f . But remember that we are now reconstructing Savage's pre-theoretic and introspective concept of preference: what is at issue is what counts as order in the mind. Arguably, it is not psychologically possible for two directly contradictory mental states such as $(f, g, >_E)$ and $(g, f, >_E)$ to be in your *conscious* mind simultaneously. But it seems entirely possible for your *constitution* to contain two such mental states.

The only reasoning rules we need to consider are instances of the following schema, which has as variables two acts and an event, denoted by f , g and E (despite the fixed meanings of f , g and E elsewhere):

Sure-thing Rule (for combining preference and indifference): $\{[(f, g, >_E), (f, g, \sim_{\Omega E}), (E, \text{partial belief})], (f, g, >_{\Omega})\}$.

Notice that to every instance of the rationality requirement STP, there corresponds an instance of the Sure-thing Rule that achieves it. It is surely uncontroversial that these rules seem right to Savage, and hence can be used by him.

How did Savage reason? Savage (1954, p. 103) tells us that, upon reflection, he has 'accepted the following way of looking at the two situations, which amounts to repeated use of the sure-thing principle'.¹³ He explains this way of looking by saying that 'one way in which Gambles 1–4 could be realised' is by a lottery with 100 tickets and the prizes specified in Table 3. (We have added the notation f , g , f' and g' for Gambles 1, 2, 3 and 4 respectively.) The claim that this lottery implements exactly the gambles in Allais's situations depends on the unproblematic assumption that each of the 100 tickets has the same subjective probability of being drawn. We will call this representation of the gambles the *state/consequence frame*.

Table 3 reveals the theoretical relationship between the Allais Paradox and Savage's P2 axiom. The table can be read as a matrix in which rows represent acts, columns represent states, and cells contain consequences. Let $\Omega = \{1, \dots, 100\}$ be the set of states, defined by which ticket number is drawn, and define the event $E = \{1, \dots, 11\}$. Table 3 tells us that in E , f agrees with f' , and g agrees with g' , while in $\Omega \setminus E$, f agrees with g , and f' agrees with g' . According to P2, $f \leq g \Leftrightarrow f' \leq g'$ in Savage's notation, and hence $(f, g, >_{\Omega}) \in C \Leftrightarrow (f', g', >_{\Omega}) \in C$ in ours. Thus, Savage's initial preferences violate P2. This conclusion gives a more precise identification of *how* those preferences contravene Savage's theory, but he still has to decide whether to reject P2 or change his preferences.

Savage's reasoning within the state/consequence frame is as follows (we have separated this quotation into numbered passages for easier reference):

Table 3. The decision situations as finally represented by Savage.

		Ticket number		
		1	2–11	12–100
Situation 1:	Gamble 1 (f)	\$500,000	\$500,000	\$500,000
	Gamble 2 (g)	\$0	\$2,500,000	\$500,000
Situation 2:	Gamble 3 (f')	\$500,000	\$500,000	\$0
	Gamble 4 (g')	\$0	\$2,500,000	\$0

[1] Now, if one of the tickets numbered from 12 through 100 is drawn, it will not matter, in either situation, which gamble I choose. [2] I therefore focus on the possibility that one of the tickets numbered from 1 through 11 will be drawn, in which case Situations 1 and 2 are exactly parallel. [3] The subsidiary decision depends in both situations on whether I would sell an outright gift of \$500,000 for a 10-to-1 chance to win \$2,500,000 – a conclusion that I think has a claim to universality, or objectivity. [4] Finally, consulting my purely personal taste, I find that I would prefer the gift of \$500,000 [5] and, accordingly, that I prefer Gamble 1 to Gamble 2 and (contrary to my initial reaction) Gamble 3 to Gamble 4. (1954, p. 103)

Table 4 represents Savage's new understanding of the decision problems, expressed in the kind of intuitive language we used in Table 2. On this new description, the two decision problems coincide; the description suppresses the difference.¹⁴ The asterisk is a placeholder for 'whatever happens, assuming one of tickets 12–100 is drawn', which (within a given situation) is the same for both options. The relevant choice is about what happens, assuming one of tickets 1–11 is drawn.

We reconstruct Savage's reasoning as a four-stage process, focusing on his reasoning about f and g ' (the acts for which he 'corrects' his preference).

Stage 1 (carry-over of automatic processing in the probability distribution frame). Responding to Allais's situations, when presented as in Table 1 (the *probability distribution frame*), Savage formed attitudes of strict preference for Gamble 1 over Gamble 2, and for Gamble 4 over Gamble 3. We are told that these mental states were formed as 'immediate express[ions]' of preference, based on 'intuitive attraction' (Savage, 1954, p. 103). We interpret this as System 1 mental processing, activated by the cues of the probability distribution framing.¹⁵ Since Savage sees f , g , f' and g' as *realisations* of Gambles 1, 2, 3 and 4 respectively, he brings these initial preferences to the new frame. At this stage, his constitution takes the form

$$C = \{(f, g, \succ_{\Omega}), (g', f', \succ_{\Omega}), \dots\},$$

where '...' stands for other attitudes, including preferences relative to other acts, beliefs, intentions, etc.

Stage 2 (automatic processing in the state/consequence frame). In 'focusing on the possibility that' one of the tickets 1–11 is drawn (event E), he engages in conditional considerations. Assuming E , f gives \$500,000 for sure, while g gives \$2,500,000 with a ten-to-one chance. Consulting his 'purely personal taste', he *finds that*, assuming E , he prefers f' to g' (passage [4]): he comes to have the attitude (f', g', \succ_E) . This attitude is not produced by conscious reasoning; it can be conceptualised as an outcome of the automatic processes of System 1, acting on the cues provided by the state/consequence framing. Assuming instead that one of the last eighty-nine tickets is drawn (event $\Omega \setminus E$), he realises that the acts f and g' yield the same payoff (\$500,000), and finds himself having a conditional indifference, formally the mental state $(f', g', \sim_{\Omega \setminus E})$. Recognising that E is a real possibility, he also forms the attitude $(E, \text{partial belief})$. As the mental states $(f', g', \sim_{\Omega \setminus E})$ and $(E, \text{partial belief})$ are so natural, we treat them as System 1 outputs. But they might instead have been generated through

Table 4. Savage's final mental representation of both decision situation.

Choose between		
Gamble 1/Gamble 3		
ticket, ..., 11 drawn		ticket 12, ..., 100 drawn
certain		*
large prize		
Gamble 1/Gamble 3		
ticket, ..., 11 drawn		ticket 12, ..., 100 drawn
likely	unlikely	*
very prize	status quo	

reasoning from more basic attitudes.¹⁶ Savage's constitution now takes the form:

$$C = \{(f, g, \succ_{\Omega}), (g', f', \succ_{\Omega}), (f', g', \succ_E), (f', g', \sim_{\Omega \setminus E}), (E, \text{partial belief}), \dots\}.$$

Stage 3 (reasoning). In our reconstruction of passage [5], Savage uses the Sure-thing Rule to reason from $\{(f, g', \succ_E), (f', g', \sim_{\Omega \setminus E}), (E, \text{partial belief})\}$ to (f, g', \succ_{Ω}) . Spelt out using marked contents, here is his reasoning:

Assuming E , rather f than g' .

Assuming not E , just as well f as g' .

Maybe E .

So, rather f than g' .

Savage's constitution now takes the form:

$$C = \{(f, g, \succ_{\Omega}), (g', f', \succ_{\Omega}), (f', g', \succ_E), (f', g', \sim_{\Omega \setminus E}), (E, \text{partial belief}), (f', g', \succ_{\Omega}), \dots\}.$$

Stages 2–3 can be repeated for f and g . Stage 2 would add (f, g, \succ_E) and $(f, g, \sim_{\Omega \setminus E})$ to Savage's constitution and 'confirm' (i.e. repeat the formation of) $(E, \text{partial belief})$. Stage 3 would confirm (f, g, \succ_{Ω}) . Savage's constitution now satisfies the relevant instances of STP.

Stage 4 (Disappearance of an attitude). If, as our reconstruction assumes, Savage's constitution still contains his original preference (g', f, \succ_{Ω}) , that constitution now violates an instance of the rationality requirement AP. Recall that in his account of his reasoning, Savage describes his reaching the conclusion (f, g', \succ_{Ω}) as 'reversing my preference between Gambles 3 and 4'; but the *reversal* of the original preference requires the extinction of a mental state as well as the formation of a replacement. In a Broomean model, reasoning cannot eliminate any mental state. By what process does (g', f, \succ_{Ω}) disappear?

Savage answers an analogous question early in *The Foundations of Statistics*. He is responding to an imagined critic who challenges the analogy between his axioms and principles of logic (see Section 2 above). Savage concedes that he cannot 'controvert' the critic; he can only offer introspection:

I would, in particular, tell him that, when it is explicitly brought to my attention that I have shown a preference for f as compared with g , for g as compared with h , and for h as compared with f , I feel uncomfortable in much the same way as when it is brought to my attention that some of my beliefs are logically contradictory. Whenever I examine such a triple of preferences on my own part, I find that it is not at all difficult to reverse one of them. In fact, I find on contemplating the three alleged preferences side by side that at least one of them is not a preference at all, at any rate not any more. (1954, p. 21)

In the Allais Paradox case, it comes to Savage's attention that he has arrived at a preference for g' over f (in Stage 1) and at a preference for f over g' (in Stage 3). It is natural to suppose that he would feel at least as uncomfortable about this anomaly as about the anomaly in the cited passage, and that the nature of the discomfort would be much the same. We therefore assume that Savage's responses to the two cases are similar. That is, on contemplating the preferences (f, g', \succ_{Ω}) and (g', f, \succ_{Ω}) side by side, he *finds that* (g', f, \succ_{Ω}) is not really or no longer a preference. The absence of this preference is not the conclusion of explicit reasoning; it is the result of an automatic mental process. As a conscious agent, Savage merely observes that a preference has disappeared.¹⁷

One might ask why, as a matter of empirical psychology, the preference that disappears is (g', f, \succ_{Ω}) and not (f, g', \succ_{Ω}) . We will return to this question in Section 5. For the moment, it is sufficient to note the implausibility of assuming that the relevant difference between the two preferences is that (f, g', \succ_{Ω}) was formed as the conclusion of a reasoning rule, while (g', f, \succ_{Ω}) was an output of automatic processing. The conclusion of a reasoning rule is no more secure than the premises from which

it was derived, and (f, g', \succ_{Ω}) was derived from premises that included the automatically-produced (f, g', \succ_E) .

We can now answer the first main question addressed by our paper: Is Savage's response to Allais's challenge vulnerable to Broome's critique? Our answer is 'No'. Taken altogether, Savage's account is remarkably consistent with Broome's account of reasoning towards rationality. We emphasize three points of convergence.

First, unlike the 'writers on rationality' that Broome criticises, Savage does not think he has finished his job when he has justified P2 as a requirement of rationality: he also makes a serious attempt to explain how someone might come to satisfy that requirement.

Second, Savage does not reason from the belief that either P2 or the Sure-thing Principle is justified. His reasoning at the crucial Stage 3 is first-order, using a reasoning rule (the Sure-thing Rule) rather than referring to a rationality requirement. And, as Broome finds important, this rule seems right to Savage, just as it will seem right to many readers of *The Foundations of Statistics*.

Third, in line with Broome's rejection of reasoning towards absences, Savage does not explain the disappearance of attitudes through explicit reasoning. In particular, he does not claim that preferences can disappear by reasoning from the belief that preferences should be consistent (which would be second-order reasoning). Instead, automatic processes extinguish preferences, or perhaps fail to refresh them and so allow their decay.

As we said in Section 1.1, Savage's response to Allais's challenge can be interpreted as a form of (auto-)preaching. Savage's choices in the lunch break violate SEUT. Savage then explains the theory to himself, and confronts himself with 'the analysis' of those choices. After receiving this explanation, he wishes to change his original choices in a way that satisfies SEUT. We have shown that all of Savage's mental operations are compatible with a Broomean model of reasoning (although not all of these are operations of reasoning). The implication is that Savage's response to Allais shows in principle how, by preaching the normativity of rationality requirements, decision theorists might induce individuals to make choices that satisfy those requirements. We emphasise *might*: we know that Savage's preaching was successful in his own case, but nothing we have said so far implies that other people would be susceptible to it. Notice also that Savage's preaching has enabled him to 'correct' violations of a closedness requirement (the Sure-thing Principle) and of a consistency requirement (asymmetry of preference) by an interplay of reasoning and automatic processes. How, if at all, he would correct a violation of a completeness requirement remains an open question.

5. Two concepts of error-correction

We now turn to our second main question: Does Savage's mode of reasoning show the legitimacy of the concept of error-correction that is used in behavioural economics?

Although we have presented a reconstruction of the mental processes by which Savage came to satisfy P2, it is not entirely clear what Savage means when he says that these processes *corrected an error* in his initial preferences. He himself admits ambivalence about this claim. Concluding his discussion of Allais's situations, he says:

It seems to me that in reversing my preference between Gambles 3 and 4 I have corrected an error. There is, of course, an important sense in which preferences, being entirely subjective, cannot be in error; but in a different, more subtle sense they can be. (1954, p. 103)

His only explanation of this 'more subtle' sense is the following example:

Let me illustrate with a simple example containing no reference to uncertainty. A man buying a car for \$2,134.56 is tempted to order it with a radio installed, which will bring the total price to \$2,228.41, feeling that the difference is trifling. But, when he reflects that, if he already had the car, he certainly would not spend \$93.85 for a radio for it, he realizes that he has made an error. (1954, p. 103)

Notice that what Savage is calling an ‘error’ is the man’s initial inclination to pay \$93.85 for the radio, not the inconsistency between his preference for buying the radio as an optional extra and his preference for not buying it as a stand-alone purchase. Similarly, the ‘error’ that Savage corrects when he reflects on Allais’s situations is that of preferring Gamble 4 to Gamble 3, not his violation of P2. Of course, Savage views that violation as *irrational* (as, presumably, he views the inconsistency in the car-buyer’s preferences); but he achieves rationality by correcting a *specific* preference. In Broome’s terms, the error in each case is an erroneous mental state, not the violation of a rationality requirement.

The anomalous preferences of the man buying the car are indeed analogous with the Allais Paradox, but this example does not take us much further in understanding what Savage means by ‘error’. Plausibly, both anomalies reflect the psychological mechanism of *diminishing sensitivity* (Kahneman & Tversky, 1979). If you are thinking about the possibility of missing out on a large fortune, the difference between probabilities of 0.90 and 0.89 seems less significant than the difference between 0.01 and zero. The probability distribution framing of the choice between Gambles 3 and 4 prompts you to compare 0.90 and 0.89; the state/consequence framing prompts you to compare 0.01 and zero. Similarly, if you are thinking about buying a car radio, the difference between spending \$2,228.41 and spending \$2,134.56 seems less significant than the difference between spending \$93.85 and spending nothing. If the radio is framed as an optional extra, you are prompted to make the first comparison; if it is framed as a stand-alone purchase, you are prompted to make the second. But in neither case does Savage explain what makes one comparison erroneous and the other correct.¹⁸

It seems clear that, in some psychologically meaningful sense of these expressions, Savage’s reflections result in his *approving of* (or *identifying with*) the preference that he calls correct, $(f, g', >_{\Omega})$, and not the one that he calls an error, $(g', f, >_{\Omega})$. This approval remains in place even when he is thinking about Allais’s situations in the probability distribution frame and even when his System 1 is reacting to that frame by producing a preference for Gamble 4 (the continuing ‘intuitive attraction’). In Broomean terms, Savage’s attitude of approval might be represented as a second-order mental state of the form $(I \text{ prefer } f \text{ to } g', \text{ approval})$.¹⁹ But nothing in Savage’s account suggests that this mental state was reached as a conclusion of reasoning.

How Savage’s attitude of approving of a preference relates to his concept of ‘correctness’ is a difficult question, which may not have a definitive answer. Remember that our concern is with Savage’s informal, introspective interpretation of ‘preference’ and not with his theoretical concept, which is explicitly based on choice. What ‘preference’ means, or should mean, in economics is a contested philosophical issue.²⁰ Under a cognitivist interpretation (favoured by Broome, 2006), a preference is a ‘better for me’ judgement, where ‘better for Savage’ is a property (objective on some interpretations, subjective on others) about which judgements can be true or false. Under an emotivist interpretation, in contrast, a preference is a relative desire. On this view, ‘Rather x than y ’ directly expresses an *attitude of* relative desire; the attitude itself is not something that can be true or false. Any of these interpretations, and a range of subtly different variants of them, could support a coherent account of what Savage means by saying that $(g', f, >_{\Omega})$ was an error and that $(f, g', >_{\Omega})$ was a correction. The cognitivist interpretation, particularly its objectivist variant, would support a more literal reading of ‘error’, but ‘making an error’ might be read simply as ‘forming a preference for which I do not have an attitude of identification’.

We have already pointed to tensions in Savage’s informal interpretation of the concept of preference (see Section 3), and have noted his ambivalence about the concept of erroneous preferences. For our present purposes, we do not need to probe more deeply into what Savage means when he refers introspectively to his own preferences. Our current question is whether the method by which Savage achieves what *he* calls ‘error-correction’ confirms the legitimacy of the methods by which behavioural welfare economists achieve what *they* call ‘error-correction’.

Recall from Section 1.2 that, in behavioural welfare economics, an individual’s latent preferences are identified by abstracting from the effects of known sources of error on observable choice behaviour. If one is to show that Savage’s response to Allais’s challenge supports the methods of

behavioural welfare economics, one needs to find some factor that was at work in Savage's formation of (g', f, \succ_{Ω}) , that can be judged to be an error according to standards that almost everyone accepts, and that Savage's reasoning corrected. Were one to find an error of this kind, one might have some reason to expect that if a person became aware of having made it, a preference that she formed to correct it would have reflective stability (i.e. she would act on that preference irrespective of how decision problems were framed).

But was there such an error in Savage's formation of (g', f, \succ_{Ω}) ? Clearly, there was no mistake of reasoning in either of Broome's senses (see Section 2): that preference was a direct output of System 1. Nor was there any failure of self-control: nothing in Savage's account suggests that, when he expressed his initial preference for Gamble 4 over Gamble 3, he acted against what *at the time* he recognised as his better judgement.

Did Savage 'incorrectly perceive the choice set', i.e. make a mistake about the nature of Gambles 3 and 4, perhaps as a result of inattention or some cognitive limitation? Notice that the probability distribution framing is explicit and transparent about the possible consequences of each gamble, expressed as amounts of money, and about the probabilities of those consequences, expressed numerically. Nothing in Savage's account suggests that he misperceived this simple information. Was there anything else he needed to know? According to Savage's SEUT, an individual's preferences between any two acts depend only on their respective consequences and probabilities. In particular, those preferences are independent of the juxtaposition of consequences that is revealed in the state/consequence framing. (For example, Table 3 is constructed in such a way that the states in which Gamble 3 gives zero are also ones in which Gamble 4 gives zero.) Indeed, it is only because Savage treats this property of independence as self-evident that he is able to treat the particular juxtaposition in Table 3 as a realisation of Allais's situations.²¹ We conclude that Savage acted on full information about each of the options in the relevant choice set.

What might be said is that the probability distribution framing fails to make salient certain kinds of relationship between objects within the choice set *and objects outside it*. In particular, it does not draw attention to the relationship between, on the one hand, an unconditional comparison between the available acts f and g' , and on the other, a comparison between the same two acts assuming E . Savage's initial failure to notice that relationship might perhaps be called 'inattention' or 'cognitive limitation'. For the sake of the argument, let us accept that this was an error by Savage.

But that is just another way of saying what has been implicit from the beginning: Savage believes that his reflective response to Allais's challenge describes a sequence of mental operations that he *could* have carried out in the Paris lunch break. Presumably, he now believes that not carrying them out was an error or oversight. If he had carried them out, he would have found that his System 1 created the conditional preference (f, g', \succ_E) *in addition to* (g', f, \succ_{Ω}) . (If he failed to notice the latter preference as an immediate effect of Allais's framing, that would be inattention too.) But that does not get us any nearer to finding what we are looking for – a factor that is an error in the sense that is recognised by behavioural welfare economists, and that contributed to the formation of the preference that Savage judges to be erroneous.

If we consider the psychological processes by which these preferences were formed, the essential difference between (f, g', \succ_E) and (g', f, \succ_{Ω}) is surely that they were formed in different frames. As Savage recognises, his formation of the preference (g', f, \succ_{Ω}) was a natural – perhaps even a reasonable – response to the cues given by the probability distribution frame. It seems that, as viewed by Savage, the mistake he made in the lunch break was to think of Allais's situations in that frame, rather than the state/consequence frame that he used in his later reflections. Ultimately, Savage endorses the preference (f, g', \succ_E) because he approves of the frame in which it is formed.

It is easy to understand why Savage privileges the state/consequence frame, particularly when he is asking himself whether his axioms are justified. He has developed a conceptual framework in which preferences are defined in relation to the acts \times states matrices of consequences that feature in that frame. In contrast, the concept of probability that is central to the probability distribution frame is not a primitive; it is a property of the SEUT representation of preferences that is

implied by his axioms. Thus, the state/consequence frame is a better representation of what, *according to Savage's theory*, is fundamental to choice under uncertainty. It is therefore neither surprising nor unreasonable that Savage feels particular identification with preferences that he has found or formed when thinking about problems in that frame.

In choosing to use the state/consequence frame in responding to Allais's challenge, Savage is deliberately affecting the cues on which his System 1 will work, in something like the way that a person with a tight budget might choose to avoid going into a high-end department store. Savage's conscious mental activity is that of 'focusing' on – directing his attention to – a frame that he has chosen to use. Thus, System 1 and System 2 are playing complementary roles. Savage is conscious of the operations of both systems and of the relationships between them, even though only System 2 is under his conscious control.

We conclude that Savage's concepts of error and correction are very different from those that behavioural welfare economists use to identify individuals' latent preferences. Thus, the fact that Savage is able to correct *what he classifies as* an error gives no support to the error-correction methods of behavioural welfare economics. To the contrary, his account describes a pair of initial preferences, i.e. (f, g, \succ_Ω) and (g', f, \succ_Ω) , that have not been influenced by *what behavioural welfare economists classify as* error, but which together violate a rationality requirement of SEUT. This is an anomaly that the methods of behavioural welfare economics cannot correct.

It is natural to ask whether behavioural welfare economists could adapt Savage's method of error-correction as an alternative way of retrieving an individual's latent preferences – perhaps through some kind of structured interviewing.²² Using such a method at the level of the individual would have the merit of not needing to appeal to standards of error that almost everyone accepts: it would use each individual's own standards. But if that method is to arrive at latent preferences that satisfy standard rationality requirements, there must be some framing of the relevant decision problems that induces preferences that satisfy those requirements and, crucially, this must be a frame *that the individual herself approves*. Imagine someone with no knowledge of decision theory who responds to Allais's situations, presented in the probability distribution frame. Like Savage, she finds she prefers Gamble 1 to Gamble 2, and Gamble 4 to Gamble 3. A behavioural economist, trying to retrieve the individual's latent preferences (or a decision theorist, acting on Gilboa's recommendation to go out and preach SEUT) guides her through the steps of Savage's reasoning. The individual might get through Stage 1 (she agrees that the state/consequence frame is a realisation of the gambles), Stage 2 (she finds the same conditional preferences and beliefs as Savage did), and Stage 3 (the Sure-thing Rule seems right to her, and she reasons to (f, g', \succ_Ω)). She gets to Stage 4 and becomes aware of having two contradictory preferences, (f, g', \succ_Ω) and (g', f, \succ_Ω) . As a result of automatic processes, one of them disappears. But if she is to satisfy P2, the preference that disappears must be (g', f, \succ_Ω) . What guarantees that? And if that preference *does* disappear, what prevents it from reappearing if she goes back to thinking in the probability distribution frame? In Savage's case, the guarantee was his approval of the state/consequence frame, but we are not entitled to assume that other people share that sense of approval.

6. Conclusion

The mental experiment we have just described has a close counterpart in a famous early hypothetical-choice experiment reported by Slovic and Tversky (1974). Participants first responded to Allais's situations, presented in the original probability distribution frame. Fifty-nine per cent of participants made Allais Paradox decisions; the others acted in consistency with P2, preferring the less risky gamble in both situations. Participants who had revealed Allais Paradox preferences were then exposed to Savage's sure-thing argument, presented by a fictional Dr S, while those whose choices were consistent with SEUT were exposed to a fictional Dr A's advocacy of Allais Paradox choices. All participants were then invited to reconsider their previous decisions. In both groups,

most participants chose to stick with their original decisions; Dr A proved to be marginally more persuasive than Dr S.

In fact, the sample size (29 participants) was too small to produce statistically significant results, but let us imagine that Slovic and Tversky's findings have been replicated in a much larger experiment. For a decision theorist who shared Gilboa's position on 'preaching', that would be evidence that, for a majority of people, the Sure-thing Principle is not a requirement of rationality. We take it that Savage (who in fact died before this experiment was reported) would not have accepted that conclusion. Recall that for Savage, the Sure-thing Principle has a status analogous with that of a rule of logic: it has a claim to 'universality' or 'objectivity'. As the author of a defence of a normative theory of rationality, he believes that he has a duty of good faith not to defend principles whose implications are contravened by preferences of *his own* that, upon careful reflection, *he* endorses. That is why it is important for him to respond to Allais. But it is entirely legitimate for him to argue in favour of principles that other people, perhaps upon equally careful reflection, reject.

In trying to convince his readers of the normativity of his theory, Savage *is* preaching. One might say that, in responding to Allais's challenge, he is preaching to himself. His reasoning provides a model of how the preaching of rationality can be immune to Broome's objections to arguments that state rationality requirements without explaining how people can come to satisfy them. But it gives no support to the claim that individuals who violate standard rationality requirements are making errors according to their own standards of correctness, and that behavioural economics has methods by which those errors can be identified and counteracted.

Notes

1. Our account of this event is based on Jallais and Pradier (2005), Moscati (2016) and Mongin (2019).
2. Dietrich et al. (2019) present a formalisation of Broome's model and examine its implications for the achievability of rationality requirements through reasoning. Dietrich and Staras (2020) and Dietrich et al. (2020) analyse its relationship to formal logic.
3. Some early psychologically-based alternatives to expected utility theory were explicitly presented as challenges to the normative status of that theory (e.g., Allais, 1979 and Loomes & Sugden, 1982), and Kahneman has sometimes resisted the 'error' interpretation of anomalies (e.g., Kahneman, 1996). There is a continuing strand of literature in which non-standard choice models are advanced with a claim of capturing aspects of rationality that standard theory ignores (e.g., Gigerenzer & Todd, 1999; Dietrich & List, 2016).
4. See the critical review of this literature by Infante et al. (2016). A similar conclusion is drawn by Rizzo and Whitman (2020: pp. 46–52).
5. Here and elsewhere, we specify a requirement by stating a condition on the constitution, with the understanding that, *formally*, the requirement is the set of constitutions satisfying that condition.
6. Under a different notion of rule, rules are schemas of 'our' rules, and 'our' rules are instances of rules. Our notion of rule is convenient here; nothing hinges on it.
7. Savage recognises that in reality there can also be 'introspective sensations of indecision or vacillation' but, on grounds of parsimony, does not represent this possibility in his model (1954, pp. 17–21).
8. If Savage had wanted to avoid *all* use of mentalistic concepts, he might have tried to justify his axioms on instrumental grounds, for example by using money pump arguments. In the light of Cubitt and Sugden's (2001) demonstration of the general invalidity of money pump arguments, we think that he made a wise choice.
9. Here we are referring to the definition of $f \leq g$ given B in the main text of *Foundations of Statistics*. In the endpapers of that book, Savage gives an alternative and more complicated definition that is independent of P2; P2 is then stated as the condition that, for all f, g and B , either $f \leq g$ given B or $g \leq f$ given B .
10. For approximate 2020 equivalents, multiply the dollar amounts by 10.
11. We treat preference and indifference as distinct attitude types, rather than using a single attitude type that corresponds with the weak preference relation \geq of standard decision theory or with Savage's formal relation \leq . The latter approach would be mathematically convenient, but its attitude type would not be psychologically natural. That would be inappropriate for a model of introspective reasoning.
12. They are not the only requirements with this status. Others include symmetry of indifference, incompatibility between preference and indifference, the Sure-thing Principle for combining preference and preference, the Sure-thing Principle for combining indifference and indifference, and various transitivity conditions for preference and indifference.

13. As Savage does not make an explicit distinction between rationality requirements and reasoning rules, it is difficult to assess whether, in 'looking at' the situations, what is to be repeatedly used is the Sure-thing Principle as a rationality requirement, a corresponding reasoning rule, or some combination of the two. We try to clarify this in our reconstruction.
14. In terms of Dietrich and List's (2016) reason-based model, in the new frame the acts f and f' have the same salient properties, as do g and g' , whereas in the old frame the salient properties differ.
15. This is our understanding of how *Savage* came to have these preferences. The existence of non-standard rational choice theories that explain the common consequence effect (e.g., Allais, 1979) suggests that other people might come to the same preferences by processes in which System 1 operations are supplemented by explicit reasoning.
16. One might specify a rationality requirement of being conditionally indifferent between acts believed to conditionally agree, and a corresponding reasoning rule by which Savage could reason from (f' agrees with g' in $\Omega \setminus E$, belief) towards the state ($f, g, \sim_{\Omega \setminus E}$). Using an attitude type of *equally strong belief*, one might specify rules by which Savage could reason from the symmetry of the 100 single-state events towards equally strong belief in each of them, and from that towards (E , partial belief).
17. On an alternative reconstruction of Savage's reasoning, his automatically-created preference for Gamble 4 over Gamble 3 is not 'refreshed' when he is thinking in the state/consequence frame, and so might disappear even before (f', g', \succ_{Ω}) is formed. Our conclusions do not depend on when the 'erroneous' preference disappears.
18. Savage's intuition seems to be that, in the car radio case, diminishing sensitivity to gains and losses of money is an expression of under-sensitivity at relatively large money amounts. But an opposite intuition (i.e. over-sensitivity at relatively small amounts) is often invoked in explaining the very high degrees of risk aversion found in experiments investigating preferences over small-stake lotteries (e.g., Rabin, 2000). On a fully subjectivist view, diminishing sensitivity is an operational concept, but under- and over-sensitivity are not.
19. Approving a mental state need not imply believing that you have it.
20. On this, see e.g. Hausman (2012) and Guala (2019).
21. Juxtaposition is relevant for some theories of choice under uncertainty, such as regret theory (Loomes & Sugden, 1982), but its irrelevance is an implication of Savage's axioms.
22. In one of the first contributions to behavioural welfare economics, Bleichrodt et al. (2001) propose interactive interviewing as the methodological gold standard for correcting 'choice inconsistencies' and thereby retrieving individuals' latent preferences.

Acknowledgements

We thank John Broome, Ben McQuillin, an editor and two anonymous referees for comments. Antonios Staras's work was supported by the University of East Anglia Faculty of Social Sciences Graduate School.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Economic and Social Research Council: [Grant Number ES/P008976/1]; H2020 European Research Council: [Grant Number 670103]; French National Research Agency [Grant Numbers ANR-17-CE26-0003, ANR-16-FRAL-0010 and ANR-17-EURE-0001].

Notes on contributors

Franz Dietrich is Research Professor at Paris School of Economics and CNRS, following an interdisciplinary career in which he received a PhD in mathematics (Oxford) and held positions as philosopher or economist at Konstanz University, Maastricht University, London School of Economics, and University of East Anglia. His research lies between economics and philosophy, primarily addressing foundational and methodological questions about decisions of individuals or groups. In particular, he works on rationality, the reason-based foundations of choice, preference change, reasoning, unawareness, moral philosophy, and the aggregation of judgments, probabilistic beliefs, or preferences.

Antonios Staras uses formal methods from economics and philosophy to address issues relating to public policy, cognitive and behavioural economics, and choice under uncertainty. He teaches economics at Cardiff Business School, Cardiff University. Before entering academia, he acted as decision scientist in various roles in industry.

Robert Sugden is Professor of Economics at the University of East Anglia in the UK. His research uses a combination of theoretical, experimental and philosophical methods to investigate issues in behavioural economics, normative economics, choice under uncertainty, the foundations of decision and game theory, the methodology of economics, and the evolution of social conventions. His current work focuses on the problem of reconciling behavioural and normative economics.

References

- Allais, M. (1953). Le comportement de l'homme rationnel devant le risque; critique des postulats et axiomes de l'école Américaine. *Econometrica*, 21(4), 503–546. <https://doi.org/10.2307/1907921>
- Allais, M. (1979). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School. In M. Allais, & O. Hagen (Eds.), *Expected utility hypotheses and the Allais Paradox* (pp. 27–145). Reidel. First published in French 1953.
- Bershears, J., Choi, J., Laibson, D., & Madrian, B. (2008). How are preferences revealed? *Journal of Public Economics*, 92(8–9), 1787–1794. <https://doi.org/10.1016/j.jpubeco.2008.04.010>
- Bleichrodt, H., Pinto-Prades, J.-L., & Wakker, P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science*, 47(11), 1498–1514. <https://doi.org/10.1287/mnsc.47.11.1498.10248>
- Broome, J. (2006). Reasoning with preferences? In S. Olsaretti (ed.), *Preferences and well-being* (Royal Institute of Philosophy Supplements, 59, pp. 183–208). Cambridge: Cambridge University Press. [doi:10.1017/CBO9780511599743.009](https://doi.org/10.1017/CBO9780511599743.009).
- Broome, J. (2013). *Rationality through reasoning*. Wiley Blackwell.
- Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Regulation for conservatives: Behavioral economics and the case for 'asymmetric paternalism'. *University of Pennsylvania Law Review*, 151(3), 1211–1254. <https://doi.org/10.2307/3312889>
- Cubitt, R., & Sugden, R. (2001). On money pumps. *Games and Economic Behavior*, 37(1), 121–160. <https://doi.org/10.1006/game.2000.0834>
- Dietrich, F., & List, C. (2016). Reason-based choice and context-dependence: An explanatory framework. *Economics and Philosophy*, 32(2), 175–229. <https://doi.org/10.1017/S0266267115000474>
- Dietrich, F., & Staras, A. (2020). Reasoning in versus about attitudes: how attitude formation is beyond logic. Available at: <http://www.franzdietch.net/Papers/DietrichStaras-ReasoningAndLogic.pdf>.
- Dietrich, F., Staras, A., & Sugden, R. (2019). A Broomean model of rationality and reasoning. *Journal of Philosophy*, 116(11), 585–614. <https://doi.org/10.5840/jphil20191161138>
- Dietrich, F., Staras, A., & Sugden, R. (2020). Beyond belief: logic in multiple attitudes. Available at: <http://www.franzdietch.net/Papers/DietrichStarasSugden-BeyondBelief.pdf>.
- Gigerenzer, G., Todd, P., & the ABC Research Group (1999). *Simple Heuristics that make us Smart*. Oxford University Press.
- Gilboa, I. (2010). Questions in decision theory. *Annual Review of Economics*, 2(1), 1–19. <https://doi.org/10.1146/annurev.economics.102308.124332>
- Guala, F. (2019). Preferences: Neither behavioural nor mental. *Economics and Philosophy*, 35(3), 383–401. <https://doi.org/10.1017/S0266267118000512>
- Hausman, D. (2012). *Preference, value, choice, and welfare*. Cambridge University Press.
- Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, 23(1), 1–25. <https://doi.org/10.1080/1350178X.2015.1070527>
- Jallais, S., & Pradier, P.-C. (2005). The Allais Paradox and its immediate consequences for expected utility theory. In P. Fontaine, & R. Leonard (Eds.), *The experiment in the history of economics* (pp. 21–42). Routledge.
- Kahneman, D. (1996). Comment [on paper by Charles Plott in same volume]. In K. Arrow, E. Colombatto, M. Perlman, & C. Schmidt (Eds.), *The rational foundations of economic behaviour* (pp. 251–254). Macmillan and International Economic Association.
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9), 697–720. <https://doi.org/10.1037/0003-066X.58.9.697>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kahneman, D., & Sunstein, C. (2006). Cognitive psychology of moral intuitions. In J.-P. Changeux, A. Damasio, W. Singer, & Y. Christen (Eds.), *Neurobiology of human values* (pp. 91–105). Springer Science and Business Media.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>
- Kőszegi, B., & Rabin, M. (2008). Choices, situations, and happiness. *Journal of Public Economics*, 92(8–9), 1821–1832. <https://doi.org/10.1016/j.jpubeco.2008.03.010>
- Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, 92(368), 805–824. <https://doi.org/10.2307/2232669>

- Manzini, P., & Mariotti, M. (2014). Welfare economics and bounded rationality: The case for model-based approaches. *Journal of Economic Methodology*, 21, 342–360. <https://doi-org.uea.idm.oclc.org/10.1080/1350178X.2014.965909>
- Mongin, P. (2019). The Allais Paradox: What it became, what it really was, what it now suggests to us. *Economics and Philosophy*, 35, 423–459. <https://doi:10.1017/S0266267118000469>
- Moscatti, I. (2016). How economists came to accept expected utility theory: The case of Samuelson and Savage. *Journal of Economic Perspectives*, 30(2), 219–236. <https://doi.org/10.1257/jep.30.2.219>
- Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68(5), 1281–1292. <https://doi.org/10.1111/1468-0262.00158>
- Rizzo, M., & Whitman, G. (2020). *Escaping paternalism: Rationality, behavioural economics and public policy*. Cambridge University Press.
- Salant, Y., & Rubinstein, A. (2008). (A, f): choice with frames. *Review of Economic Studies*, 75(4), 1287–1296. <https://doi.org/10.1111/j.1467-937X.2008.00510.x>
- Savage, L. (1954). *The Foundations of Statistics*. Wiley.
- Slovic, P., & Tversky, A. (1974). Who accepts Savage's axiom? *Behavioral Science*, 19(6), 368–373. <https://doi.org/10.1002/bs.3830190603>
- Thaler, R. (2015). *Misbehaving: How economics became behavioural*. Allen Lane.
- Thaler, R., & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Wason, P., & Evans, J. (1974). Dual processes in reasoning? *Cognition*, 3(2), 141–154. [https://doi.org/10.1016/0010-0277\(74\)90017-1](https://doi.org/10.1016/0010-0277(74)90017-1)