Preference purification and the inner rational agent:
a critique of the conventional wisdom of behavioural welfare economics

Gerardo Infante (School of Economics, University of East Anglia)

Guilhem Lecouteux (Département d'économie, École Polytechnique)

Robert Sugden (School of Economics, University of East Anglia)

Abstract

Neoclassical economics assumes that individuals have stable and context-independent preferences, and uses preference-satisfaction as a normative criterion. By calling this assumption into question, behavioural findings cause fundamental problems for normative economics. A common response to these problems is to treat deviations from conventional rational-choice theory as mistakes, and to try to reconstruct the preferences that individuals would have acted on, had they reasoned correctly. We argue that this *preference purification* approach implicitly uses a dualistic model of the human being, in which an inner rational agent is trapped in an outer psychological shell. This model is psychologically and philosophically problematic.

Acknowledgements

An early version of this paper was presented at a conference of the Network for Integrated Behavioural Science at the University of Warwick in September 2014. We thank participants at that conference for comments. Infante's and Sugden's work was supported by the Economic and Social Research Council through that Network (grant reference ES/K002201/1).

Keywords: preference purification; inner rational agent; behavioural welfare economics; libertarian paternalism; context-dependent preferences

JEL codes: B41 (economic methodology); D03 (behavioural microeconomics: underlying principles); D60 (welfare economics: general).

In neoclassical economics, it is standard practice to assume that individuals have stable and context-independent preferences over all economically relevant outcomes, and to use the satisfaction of those preferences as the primary normative criterion. By calling this assumption into question, the findings of behavioural economics are causing fundamental problems for normative economics. In this paper, we critically evaluate a response to these problems that has been advocated by many prominent behavioural economists, and that has recently been endorsed by Hausman (2012) in a philosophical enquiry into how the concepts of preference, value, choice and welfare are (and ought to be) used in economics. Following Hausman (p. 102), we will call this approach 'preference purification'. The essential idea is that when an individual's decisions are inconsistent with defensible assumptions about rational choice, those decisions can be treated as mistakes. The task for welfare economics is then to reconstruct the preferences that the individual would have acted on, had her reasoning not been distorted by whatever psychological mechanisms were responsible for the mistakes, and to use the satisfaction of these reconstructed preferences as a normative criterion.

We will argue that this approach implicitly uses a dualistic model of the human being, in which an inner rational agent is trapped inside a psychological shell. The inner agent is pictured as the locus of the identity of the human being and as the source of normative authority about its interests and goals. There is no attempt to represent the psychology of this agent; its rationality is simply taken as given. The psychological mechanisms that induce deviations from supposedly rational choice are treated as properties of the outer shell that can prevent the inner agent from achieving its objectives. Whether viewed in the perspective of psychology or of philosophy, this model is problematic.

We will begin by describing some of the context-dependent features of real decision-making behaviour that cause problems for conventional welfare economics (Section 1). We will explain the preference purification approach that behavioural welfare economists have used to try to resolve these problems (Section 2), and consider Hausman's endorsement of this approach (Section 3). Drawing on Hausman and Welch's (2010) attempt to define a concept of autonomy that is appropriate for behavioural agents, we will explain the sense in which the preference purification approach presupposes the model of the inner rational agent (Section 4). We will argue that the idea that context-dependent choices are caused by errors of reasoning is fundamentally misconceived (Section 5). Finally, we will offer a conjecture

about why behavioural economists have been attracted by the model of the inner rational agent (Section 6).

1. Background

Our main focus will be on a class of cases that feature prominently in discussions about the normative significance of behavioural findings. These are cases in which a person's preferences, choices or judgements are strongly affected by factors that work through well-understood psychological mechanisms but seem to have little or no relevance to that person's well-being, interests or goals. Although there is a clear sense in which the choices made (or preferences revealed, or judgements expressed) by the person in different contexts are inconsistent *with one another*, it is not at all obvious *which* (if any) of these choices is correct – or even how 'correctness' should be defined.

Here is a typical example. In an experiment reported by Kahneman, Knetsch and Thaler (1990: 1338–1339), student subjects reported their valuations for coffee mugs. Subjects were randomly assigned to experimental treatments. In one treatment, each subject was asked to consider each of a range of amounts of money, and to say whether she would choose to have a mug or the money. In another treatment, each subject was first given the mug, free of charge, and then asked whether she would choose to sell it back to the experimenters at each of a range of prices (the same range of money amounts as in the first treatment). Notice that, defined in terms of what a subject can take away from the experiment, the problems faced by the two sets of subjects are exactly the same: the only difference is whether the problems are framed as *choosing* between the mug and money, or as *selling* the mug. However, the median valuation of the mug in the selling treatment ($7.12) was more than double that in the choosing treatment ($3.12). This effect can be explained by the hypothesis that losses have greater psychological salience than equal and opposite gains. (In the choosing treatment, subjects are thinking about *gaining* the mug, while in the selling treatment, they are thinking about *losing* it.) It would be very difficult to argue that the difference between being told that you have been given a coffee mug and being told that you can choose to be given one is a good reason for a two-fold difference in your valuation of the mug, and in this sense the effect seems irrational; but that does not answer the question of whether $7.12 is an irrationally high valuation or whether $3.12 is an irrationally low one.

Here is another example. Read and van Leeuwen (1998) report a field experiment in which workers made choices between free snacks which would be delivered at a designated

time a week later. The menu from which subjects could choose contained healthy options (e.g. apples) and unhealthy ones (e.g. Mars bars). There were four treatments, defined by two different times of day – 'after lunch time' and 'in the late afternoon' – at which the choice was made and (independently) at which the snack would be delivered. The background assumption was that most workers would be hungrier at the later time. Read and van Leeuwen found that, holding constant the time of delivery, subjects were more likely to choose unhealthy snacks if they made the choice in the late afternoon. In broad terms, the psychological mechanism behind this result is easy to understand. The hungrier you feel, the more attention you give to cues that are directed towards the satisfaction of hunger, and the more vividly you can imagine experiencing feelings of hunger in other situations. Thus, the hunger-satisfying properties of the Mars bar are perceived more vividly in the late afternoon, irrespective of when it will actually be eaten. Given the familiarity of the snack options and the predictability of daily fluctuations in hunger and satiation, it would be implausible to claim that differences in the time of day at which the decision is made provide good reasons for different choices about what to eat at a given time seven days later. In this sense, the context-dependent preferences revealed in the experiment seem irrational. But that does not answer the question of whether, in any given situation, it is more rational to choose an apple or a Mars bar.

Our third example concerns a less obvious principle of consistency. It is the version of the Allais Paradox discussed by Savage (1954, pp. 101–103). Respondents are asked to imagine two different situations, in each of which there is a choice between two gambles. In Situation 1, the choice is between Gamble 1, which gives $500,000 with probability 1, and Gamble 2, which gives $2,500,000 with probability 0.1, $500,000 with probability 0.89, and nothing with probability 0.01. In Situation 2, the choice is between Gamble 3, which gives $500,000 with probability 0.11 and nothing with probability 0.89, and Gamble 4, which gives $2,500,000 with probability 0.1 and nothing with probability 0.9. Many people report strict preferences for Gamble 1 in Situation 1 and for Gamble 4 in Situation 2. According to the axioms of expected utility theory, a person with consistent preferences would *either* prefer Gambles 1 and 3 *or* prefer Gambles 2 and 4. But the theory does not say which of those two patterns of preference is more rational.

In this paper we will focus on cases, like those we have just discussed, in which *choices* or *preferences* are allegedly inconsistent. However, it is important to keep in mind that *judgements* can also be systematically context-dependent in ways which do not seem to

be supported by defensible reasons, and that the question of which of the mutually inconsistent judgements is correct can be no easier to answer than analogous questions about choices or preferences. For example, people's judgements about their own happiness are subject to 'focusing illusions' that seem to result from mechanisms similar to those involved in the choice between future snacks in Read and van Leeuwen's experiment. When people are trying to judge their overall satisfaction with their lives, the implicit weights they give to different aspects of life can depend on what is currently the focus of their attention (Schkade and Kahneman, 1998) – an effect summed up in Kahneman's (2011: 402) maxim: 'Nothing in life is an important as you think it is when you are thinking about it'. Since most of the happiness data that economists and psychologists use are generated from self-reported judgements, one should not assume that the problem of context-dependent preferences can be resolved simply by defining 'true preferences' in terms of happiness.

We recognise that there are some cases of context-dependent choice for which the definition of a person's true preferences or best interests is fairly uncontroversial. For example, in some retail energy markets, competing suppliers offer exactly the same product, priced according to different tariffs. It seems unexceptionable to assume that, for any given quantity bought, consumers have an underlying preference for paying less rather than more. In fact, when tariffs are complex, consumers often fail to buy from the supplier offering the lowest final price (Wilson and Waddams Price, 2010). Representing such choices as mistakes, defined relative to 'true' preferences for low prices, may be a reasonable modelling strategy. In this case, however, the assumption that is taken to be uncontroversial in defining mistakes equates the consumer's *subjective* ranking of options (alternative tariffs) with an *objective* ranking (in inverse order of their prices) that is independent of the consumer's perceptions or judgements. There is no obvious analogue to this objective ranking in cases such as the choice between buying or not buying a coffee mug, or between eating an apple and eating a Mars bar.

The distinction between objective and subjective rankings is important because of the way in which the idea of mistakes is used in behavioural welfare economics. A recurring theme in this literature is that the findings of behavioural economics justify policies which 'nudge' individuals towards those choices that are in their best interests (e.g. Camerer et al., 2003; Sunstein and Thaler, 2003 [henceforth 'ST']; Thaler and Sunstein, 2008 [henceforth 'TS']). The element of paternalism in these proposals can be made more palatable by suggesting, not only that their aim is to increase the welfare of the targeted individuals, but

also that welfare is being measured *according to those individuals' own judgements*, and that the choices that individuals are being nudged away from would be *mistakes*. These suggestions are often expressed through the idea that nudges *help* individuals to make what, on reflection, they themselves would recognise as better choices. For example, asking the reader to consider a problem of choosing between a large number of prescription drug plans, Thaler and Sunstein (2008, p. 10) say: '[Y]ou might benefit from a little help. So long as people are not choosing perfectly, some changes in the choice architecture could make their lives go better (as judged by their own preferences…)'.[1] Contrast this notion of helping people to avoid mistakes with the more overt paternalism of a parent who limits a two-year-old child's consumption of chocolate in the interests of a balanced diet. The parent believes her action promotes the child's welfare, but the child's wish to eat chocolate is not a mistake. As a real desire for an experience that really is pleasurable, it makes good sense in terms of all the reasons that the child is capable of understanding. If context-dependent choices can be represented as mistakes, the relationship between a paternalistic policy-maker and a targeted individual looks less like that between a benevolent guardian and an incompetent ward.

To avoid misunderstanding, however, we must make clear that this paper is not primarily concerned with whether (or how far) public policy should be paternalistic. It is possible to investigate questions about individual welfare without presupposing that governments ought to adopt whatever policies can be shown to maximise well-being. Although many advocates of preference purification present it as a technique that can be used in designing paternalistic policies, this linkage is not universal. In particular, Hausman (2012) endorses preference purification as a tool for the measurement of welfare in applied economics, but has serious reservations about the use of nudges as a policy tool. By bracketing out the question of what governments ought to do with welfare measurements, we are able to evaluate Hausman's philosophical arguments for preference purification.

We also bracket out questions about the use of nudges for non-paternalistic purposes. A significant part of the nudge literature is directed at using behavioural insights to induce 'behaviour change' in situations in which the targeted individuals do not seem to be making mistakes in satisfying their own preferences or in promoting their own welfare: they are simply frustrating the achievement of some public policy objective. For example, TS's catalogue of emulation-worthy policies includes nudges designed to reduce littering, to

[1] TS repeatedly describe nudges as 'helping' the individuals at whom they are directed. Looking only at their first chapter, one finds this use of 'helping' on pp. 6, 7, 9, 10, 11 and 14.

increase registration in organ donation programmes, and (through naming and shaming polluting firms) to reduce the release of potentially hazardous chemicals into the environment (pp. 60, 175–182, 190–191). The UK Behavioural Insights Team prides itself on having designed nudges which make people more likely to pay their tax bills on time, to the benefit of the public finances (Halpern and Nesterak, 2014). Discussion of the legitimacy of such policies has focused on issues of transparency and democratic accountability (Hansen and Jespersen, 2013; Sunstein, 2014). In contrast, our concern is with the definition and measurement of welfare.

2. Behavioural welfare economics and preference purification

Since Sunstein and Thaler have been particularly influential in the development of behavioural welfare economics, we being by looking at the role of preference purification in their arguments. Their original paper (ST) sets out a manifesto for *libertarian paternalism*.[2] Their later book *Nudge* (TS) extends and popularises the ideas in ST.

One of Sunstein and Thaler's key claims is that the findings of behavioural economics make paternalism unavoidable: the anti-paternalist position is 'incoherent', a 'nonstarter'. In both works, this claim is developed in relation to a now-familiar cafeteria example. The premise is that customers' choices between alternative food items are influenced by the prominence with which those items are displayed on the cafeteria counter. Knowing that some items are healthier than others, the cafeteria director has to choose the relative prominence with which different items are displayed. ST consider two apparently reasonable strategies that the director might adopt: she could 'make the choices that she thinks would make the customers best off, all things considered' or she could 'give consumers what she thinks they would choose on their own'. We are told that the second option is 'what anti-paternalists would favor', but that the anti-paternalist argument for this option is incoherent. By assumption, the customers

> lack well-formed preferences, in the sense of preferences that are firmly held and preexist the director's own choices about how to order the relevant items [along the

[2] Contemporaneously with ST, Camerer et al. (2003) advocated 'asymmetric paternalism' as a normative response to the findings of behavioural economics. There are close similarities between the two proposals. Asymmetric paternalism is presented as a way of helping boundedly rational individuals to avoid 'decision-making errors' that 'lead people not to behave in their own best interests' (pp. 1211–1212). However, Camerer et al. give even less guidance than ST about how individuals' interests are defined or how they can be identified.

> counter]. If the arrangement of the alternatives has a significant effect on the selections of the customers make, then their true 'preferences' do not formally exist.[3]

Sunstein and Thaler conclude that the first strategy, despite being paternalistic, is the only reasonable option for a well-intentioned director (ST: 1164–1165, 1182; see also TS: 1–3).

Notice that, as in the general class of problems described in Section 1, the choices of the cafeteria customers are context-dependent in a way that has a psychological explanation (more prominently-displayed items are more likely to engage attention) but does not seem relevant to customers' interests or goals. Such cases are central to Sunstein and Thaler's argumentative strategy. The key innovation of libertarian paternalism is the idea that individuals' choices from given sets of (objectively defined) options can be influenced by interventions which affect only the (subjectively perceived) framing of the decision problem. Such nudges can work only in cases in which choices are context-dependent.

Notice also that the cafeteria problem is presented as a problem *for the cafeteria director*. The director is understood as someone who acts on her own authority and responsibility, but with the objective of benefiting her customers. Sunstein and Thaler describe this role as that of a 'planner' (in ST) or 'choice architect' (in TS). The idea that normative recommendations are addressed to a benevolent planner is a common device in welfare economics, and leads naturally to the further idea that those recommendations should be directed at increasing the well-being of the individuals for whom the planner is planning. In TS, this idea is given a slightly different twist: Sunstein and Thaler say that their recommendations are designed to 'make choosers better off, *as judged by themselves*' (TS: 5; italics in original). The italicised clause recurs with minor variations throughout TS (e.g. pp. 10, 12, 80). The implication, we take it, is that although the planner acts on her own responsibility, she tries to respect each individual's subjective judgements about what makes him better off.

Sunstein and Thaler's approach to normative economics requires that the planner can reconstruct each individual's judgements about his own well-being, even though these judgements are not always revealed in his choices. But how, at the conceptual level, are we

---

[3] We take it that, in this passage, ST are using 'preference' in the sense that it is used in conventional economic theory – that is, as a binary relation over potential objects of choice that is consistently revealed in an individual's decisions. In this sense, the cafeteria customer does not have well-defined ('true') preferences over food items. In Section 1 above, we followed a common practice in behavioural economics by using the concept of 'true preference' is a different sense – to refer to the preferences on which (it is supposed) an individual would act in the absence of errors.

to understand these judgements? And how is the planner to reconstruct them? The closest that Sunstein and Thaler come to addressing these questions systematically is in their discussion of decision-making errors.

Immediately after presenting the principle of trying to make choosers 'better off, *as judged by themselves*', TS undertake to show that

> in many cases, 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. (TS: 5)

The corresponding passage in ST uses almost the same characterisation of decisions that would not have been made if individuals had been fully rational, and refers to these as 'inferior decisions in terms of their [i.e. the individuals'] own welfare' (ST: 1162). The implication is that, for Sunstein and Thaler, the criterion of individual well-being is given by the preferences that the relevant individual would have revealed, had his decision-making not been affected by *reasoning imperfections* – that is, limitations of attention, information, cognitive ability or self-control. So the task for the planner is to try to reconstruct individuals' underlying or *latent* preferences by simulating what they would have chosen, had they not been subject to reasoning imperfections.[4] This is *preference purification*.

Notice that preference purification cannot provide the welfare criterion that Sunstein and Thaler need unless latent preferences are context-independent. The context-dependence of revealed preferences, with the supposed implication that paternalism is unavoidable, provides the starting point for Sunstein and Thaler's argument for libertarian paternalism. But if the choice architect's decision criterion turned out to be context-dependent too, that argument would be fatally undermined. The assumption that latent preferences are context-independent is implicit in Sunstein and Thaler's arguments, but is never defended. One of their favourite rhetorical strategies is to characterise their opponents as maintaining that human beings are not subject to reasoning imperfections – that human beings can 'think like Albert Einstein, store as much memory as IBM's Big Blue, and exercise the willpower of Mahatma Gandhi'. The reader is invited to agree with the hardly controversial proposition that 'the folks we know are not like that', and encouraged to infer that *this must be why* ordinary folks' choices have the apparently irrational patterns that behavioural economics and

[4] The term 'latent' is borrowed from Kahneman's (1996) critique of Plott's (1996) 'discovered preference hypothesis'. Kahneman characterises Plott's approach as attributing latent rationality to economic agents whose behaviour contravenes neoclassical theory.

psychologists have discovered (TS: 6–8). But this inference is not as obviously valid as it may seem. We will return to this issue later, but to lay a foundation for later arguments we invite the reader to think about the following question. Imagine a being – let us call him SuperReasoner – who has the intelligence of Einstein, the memory of Big Blue, and the self-control of Gandhi. Imagine in addition (since this is also part of Sunstein and Thaler's characterisation of perfect reasoning) that SuperReasoner's capacious memory contains every item of information that can be extracted from any existing publication or database. Otherwise, however, SuperReasoner is just like some ordinary human, whom we will call Joe. If Joe were in Sunstein and Thaler's cafeteria, his choices between food items would be influenced by the prominence of their displays. Now imagine taking SuperReasoner into the cafeteria. Would the probability of his choosing cream cake be independent of the position of cream cake on the counter?

Bernheim and Rangel (2007, 2009) propose an approach to behavioural welfare economics that is similar to preference purification. Presuming it to be self-evident that welfare economics is addressed to 'the planner', they characterise 'standard welfare economics' as 'instruct[ing] the planner to respect the choices an individual would make for himself' (2007: 464). Their objective is to extend this form of welfare economics to cases in which choices are context-dependent. The key concept in their theoretical framework is a *generalised choice situation* (GCS) for a given individual, consisting of a set of 'objects' from which the individual must choose one, and a set of 'ancillary conditions'. Ancillary conditions are properties of the choice environment that may affect behaviour but which the planner treats as normatively irrelevant. (Applying this conceptual scheme to the cafeteria, food items are objects while ways of displaying them are ancillary conditions.) The individual's choice behaviour is represented by a correspondence which, for each GCS, picks out the subset of objects that the individual is willing to choose. Bernheim and Rangel's first line of approach is to propose a criterion that respects the individual's revealed preferences over pairs of objects if those preferences are not affected by changes in ancillary conditions, and instructs the planner 'to live with whatever ambiguity remains' (2009: 53). They then suggest that this rather unhelpful criterion might be given more bite by the deletion of 'suspect' GCSs. A GCS is deemed to be suspect if its ancillary conditions induce impairments in the individual's ability to attend to or process information, or to implement desired courses of action. In effect, this approach purifies choice data by eliminating any choices that were made when the individual's reasoning was impaired. Considering only the

purified data, it then uses the satisfaction of context-independent revealed preferences as the normative criterion. Notice that this approach yields welfare rankings only for those pairs of objects for which revealed preferences, after purification, are context-independent.

A different way of using the idea of purification is to begin by assuming the existence of context-independent latent preferences, and to propose some specific model of the psychological processes that intervene between those preferences and actual choices. Given such a model, one can then investigate how far and under what assumptions latent preferences can be reconstructed from observations of choices. Salant and Rubinstein (2008) follow this approach within a general theoretical framework similar to Bernheim and Rangel's. They define an *extended choice problem* for an individual as a pair (A, f) where A is a set of mutually exclusive and exhaustive alternative objects of choice, and f is a 'frame'. The individual's decisions are determined by the interaction of her frame-independent 'underlying preferences' with decision-making heuristics that are activated by, and conditional on, the frame (p. 1288). Like Sunstein and Thaler, they imagine a 'social planner' who chooses the frame with the aim that the individual's choice should be consistent with her underlying preferences (p. 1294).

This approach is often used to derive normative implications from behavioural models. The model of 'salience and consumer choice' presented by Bordalo, Gennaioli and Shleifer (2013, henceforth 'BGS') is a good example. The psychological intuition behind this model is that choice options ('goods') can be described as bundles of characteristics, and that when a consumer evaluates a good, she gives most attention, other things being equal, to those attributes of that good that are perceived as most 'salient' (that is, as having values that are most different, positively or negatively, from the average values of all the goods in the choice set). BGS assume that a 'rational' consumer would maximise a linear utility function in which each attribute has a constant utility weight; by implication, these weights represent the consumer's true subjective valuations of the attributes. A non-rational consumer maximises a function in which the attribute weights are 'distorted' by salience, and so may 'undervalue' or 'overvalue' a good, depending on how its attributes compare with the corresponding averages. In our terminology, BGS are treating the rational utility function as representing the latent preferences of the non-rational consumer. The properties of the model ensure that, given sufficient observations of the choices of a non-rational consumer, her latent preferences can be recovered.

This methodology is developed with a more applied emphasis by Bleichrodt, Pinto-Prades and Wakker (2001; henceforth 'BPW') and Li, Li and Wakker (2012). These authors are primarily concerned with cases in which a professional specialist has to make a decision in the best interests of a client. For example, consider a physician who has to choose between alternative medical treatments for an unconscious patient. The physician has access to data from a stated-preference survey in which the patient made various hypothetical choices between alternative probability distributions over health states. However, these responses are not fully consistent with one another, given the background assumption that 'the right normative model for decision under uncertainty' is expected utility theory – an assumption to which BPW are committed. According to BPW, such inconsistencies in stated-preference responses 'designate deficiencies in our measurement instruments that, even if the best currently available, do not tap perfectly into the clients' values'. The problems resulting from inconsistencies in stated preferences can be mitigated if those preferences are elicited in face-to-face interviews in which the client is asked to reconsider inconsistent choices (pp. 1498–1500, 1510). Notice the implicit assumption that the client has (or can be guided to form) preferences that are consistent with one another and with expected utility theory; the use of interviews is presented as a method of purifying preferences by the elimination of error.

But what if the physician has to make do with the patient's inconsistent survey responses? The real novelty of BPW's approach is their proposal of 'a quantitative manner for correcting biases in decision under risk and uncertainty when these cannot be avoided' (p. 1499). BPW use cumulative prospect theory (Tversky and Kahneman, 1992) as the *descriptive* model of choice while retaining expected utility theory as the *normative* model. There are two main differences between these models. First, cumulative prospect theory uses a *probability weighting function* to transform objective probabilities into their subjective counterparts; this transformation can be interpreted as taking account of psychological biases in the processing of probability information. Second, it has a *loss aversion* parameter which can be interpreted as picking up a bias induced by the framing of decision problems. Given these interpretations, an expected utility model of preferences can be constructed from an empirically estimated prospect theory model by replacing the estimated probability weighting and loss aversion parameters with the 'unbiased' values implied by expected utility theory. BPW propose that the patient's stated preferences should be used to estimate a prospect theory model, and that the 'corrected' expected utility model should be used to make choices on behalf of the patient. This is an econometric form of preference purification.

A somewhat similar methodology is proposed by Kőszegi and Rabin (2007, 2008), who frame the problem as that of making inferences about individuals' preferences from observations of their choices while recognising that the reasoning that led to these choice may have involved mistakes. Their main examples are of choice under uncertainty. Their approach is to infer an individual's subjective beliefs from the choices he makes between gambles with money outcomes, on the assumption that he prefers more money to less (contingent on any given state) and prefers higher probabilities of preferred outcomes to lower probabilities. If subjective beliefs, elicited in this way, do not coincide with objective relative frequencies, a 'revealed mistake in beliefs' is deemed to have occurred. The individual's preferences are then purified by working out what he would have chosen, had he acted on correct beliefs.

The work we have reviewed in this Section can be understood as belonging to a common programme for reconciling normative economics with behavioural findings. This programme of *behavioural welfare economics* takes the objective of normative economics to be the measurement of the effects of economic policies on individual well-being, as assessed from the viewpoint of a social planner or entrusted professional (such as a physician, dietician or 'choice architect') who wishes to respect individuals' judgements about their own well-being. It treats cases in which an individual's choices depend on 'irrelevant' properties of framing as errors, 'error' being defined relative to the latent preferences that the individual would have revealed if not subject to reasoning imperfections. Latent preferences are assumed to satisfy conventional principles of rational consistency – in particular, context-independence. The satisfaction of latent preferences is taken as the normative criterion.

Implicit in this programme, as we understand it, is the idea that latent preference is a subjective concept. By this we mean that latent preferences are judgements or perceptions that are formed within the minds of individual human beings; they do not correspond directly with objective properties of the external world. Sunstein and Thaler appeal to this notion of subjectivity when they repeatedly insist that their aim is to make people better off *as judged by themselves*. Quite apart from issues of rhetorical strategy, there is a fundamental reason for thinking that the preference purification approach presupposes a subjective interpretation of latent preferences. Consider the implications of the contrary position, that behavioural welfare economics uses a criterion of (supposedly) objective well-being. What then would be the point of taking the circuitous route of reformulating that criterion as the satisfaction of latent preferences, defining a person's latent preferences in terms of the hypothetical choices

that he would make in the absence of reasoning imperfections, and then postulating that such choices would maximise objective well-being? But if well-being is a subjective concept, preference purification can be defended as a way of correcting errors in an individual's reasoning while respecting his judgements about his own well-being.

If the purification approach is to work, latent preferences must be coherent. But if latent preference is a subjective concept, the coherence properties that are attributed to them cannot be explained by the hypothesis that latent preferences map some objective concept that already has those properties. So the preference purification approach, as applied to any given individual, must presuppose that the individual has potential access to some mode of *latent reasoning* that generates subjective preferences that satisfy conventional principles of rational consistency. However, the writers we have considered so far do not explain what that mode of reasoning is, or how it generates coherent preferences. All they tell us is that it is free of the 'imperfections' that behavioural economists and cognitive psychologists have identified in actual human reasoning. This limitation of the preference purification approach perhaps stems from the structure of the standard theory of rational choice. That theory is formulated in terms of axioms of consistency among preferences, and between preferences and choices; it does not try to explain the reasoning by which individuals construct their preferences. In an analysis which uses this conceptual framework, latent reasoning is a black box. One of the interesting features of Hausman's defence of preference purification is that it looks inside this box.

3. Hausman on preference purification

Hausman (2012) discusses preference purification as part of a larger analysis of the economic concepts of preference and welfare. He says that this analysis 'clarifies and for the most part defends' the everyday practice of economics, while challenging some of the ideas that economists use when philosophizing about their work (p. i).

Hausman begins by trying to find a coherent interpretation of the concept of preference, as that is standardly used in positive economics. He proposes the following definition: 'To say that Jill prefers *x* to *y* is to say that when Jill has thought about everything she takes to bear on how much she values *x* and *y*, Jill ranks *x* above *y*' (p. 34). Thus, a preference is *comparative* (*x* is *ranked* above *y*); the comparison is in terms of *value*; the valuation is *subjective* ('how much *she* values …'); and it takes account of the *totality* of factors that the individual thinks relevant to the comparison ('*everything* she takes to bear on

…'). In short, a preference is a 'total subjective comparative evaluation'. Hausman claims that this definition 'matches most of current practice' in economics, and urges economists to reserve the word 'preference' for this usage (pp. 34–35).

That the economic concept of preference is comparative and subjective seems uncontroversial. That it is also total is implicit in a fundamental feature of the role of preferences in economics – that preferences determine choices. (A person's choices can be influenced by any factors that she takes to be relevant. So if preferences determine choices, preferences must take account of any such factors too.) But Hausman's claim that preferences are *evaluations* is tied in with his reason-based understanding of choice. Since this claim turns out to be important for Hausman's defence of preference purification, we need to explain what he means by it.

Hausman interprets preferences as products of reasoning, and as premises that can be used in further reasoning about what to choose. It is, he says, a misconception to think that preferences are 'arbitrary matters of taste, not subject to rational consideration' (p. 18); they are 'more like judgments than feelings' (p. 135). He interprets the economic theory of choice as a theory of rational deliberation, in which individuals try to answer the question 'What do I have most reason to do?' (p. 5). He maintains that, in using a theory of rational choice, economics is committed to the claim that its explanations of an individual's choices are expressed in terms of reasons that *justify* those choices. Thus:

> [E]conomists regard ordinal utility theory as both a fragment of a positive theory that explains and predicts choices and as a fragment of a theory of rational choice that specifies conditions that preferences must satisfy in order to justify choices. (p. 20)

In arguing that this interpretation of 'preference' is faithful to the practice of economics, Hausman (pp. 19–20) considers how the axioms of choice theory might be justified as properties of sound reasoning about choice. He agrees with Broome (1991) that the logic of total comparative evaluation requires transitivity. (If, all things considered, x is more valuable than y and y is more valuable than z, then necessarily, x is more valuable than z.) While not claiming that the completeness axiom is logically required in the same way, Hausman points out that if an individual's choices are always (i.e. given any feasible set) to be determined by preferences, preferences must be complete. Thus, completeness is 'a boundary condition on rational choice'. The implicit axiom that preferences are context-independent excludes 'factors that ought to be irrelevant' for total comparative evaluations.

Thus, if a rational choice is one that is justified by sound and relevant reasons, and if reasons are to take the form of total comparative evaluations of the feasible options, rational choice is possible in general only if those evaluations are complete, transitive and context-independent. In this sense, Hausman's interpretation of 'preference' offers an explanation of why the axioms of choice theory are treated as principles *of rationality*.

Having settled on a definition of 'preference', Hausman goes on to consider the role of preferences in welfare economics. He takes it as self-evident that welfare economics is concerned with individual well-being, assessed from some neutral viewpoint. He does not specify explicitly who is the addressee of welfare economics, but his intermittent references to 'legislators', 'policy makers' and 'policy analysts' (e.g. pp. 89, 93, 95–97, 100–101) imply that, in the language of economics, the addressee he has in mind is a social planner who is seeking to promote well-being. In these respects, Hausman's approach to normative economics is aligned with that of behavioural welfare economics.

One of Hausman's central claims is that 'the satisfaction of preferences – even when preferences are informed, rational, and generally spruced-up – does not constitute well-being' (p. 77). Hausman sets out to show that preference-satisfaction theories of well-being – that is, theories that treat preference satisfaction as a (or even the only) constituent of well-being – are 'mistaken' and 'untenable' (pp. 86, 88). In line with his claim to defend the 'most part' of the practice of economics, Hausman allows that, in fact, satisfying preferences usually contributes to well-being. But this is because, in arriving at total evaluations of the options from which they can choose, individuals are in most cases strongly influenced by reasonable beliefs about how each option would affect their well-being. In such cases, preferences provide reliable *information about* well-being, but preference-satisfaction still does not *constitute* well-being.

However, Hausman (pp. 81–83) points to various cases in which, he claims, the satisfaction of preferences may *not* promote well-being. One such case is where individuals' evaluations of options are based on beliefs that are in fact false. Another is where individuals consciously choose to act contrary to self-interest. For our purposes, the most relevant case is where 'preferences are the result of … problematic psychological mechanisms'. Following the practice of behavioural welfare economics, Hausman treats these mechanisms as inducing *mistakes*:

> Contemporary psychology has identified contexts in which people are likely to make mistakes, and policy analysts can use these findings to help decide whether

> to take people's preferences as guides to their welfare. One advantage of understanding that preferences are total comparative evaluations is that economists and regulators can make clear sense of people's preferences being *mistaken*. (p. 100, italics in original)

Expanding on this idea in relation to cost-benefit analysis, Hausman says that in deciding whether to use preferences as indicators of well-being, one should ask whether the context in which the preferences are revealed is one in which preferences are 'undistorted'. Having defined the 'net benefit' of a policy of the excess of gainers' willingness to pay over losers' willingness to accept (p. 93), he says:

> The cost-benefit analyst should avoid relying on net benefits when preferences are distorted by decision-making flaws because the flaws provide a good reason to doubt that such preferences are a good guide to the individual's welfare. Examples of such flaws include overconfidence, exaggerated optimism, status quo bias, inertia, inattention, myopia, conformity, akrasia, and addiction. (p. 100)

So what *should* cost-benefit analysts rely on? Hausman's answer is that '[t]he best economists can do when they recognize flaws in people's deliberative capacities is to minimize their influence'. More specifically:

> [W]hen preferences are self-interested, well-informed, and undistorted … it is sensible for those seeking to promote welfare to employ methods of appraising policies such as cost-benefit analysis that rely on information concerning preference satisfaction. When these conditions are not met, it makes sense to take steps to purify people's preferences of mistake and distortion so as to widen the domain in which these conditions are met and to attempt to measure expected benefit rather than preferences. (p. 102).

So Hausman's proposal for dealing with preference inconsistencies is essentially the same as that of behavioural welfare economics: preference purification.

Recall Hausman's argument that if preferences are to provide reasons for choice, they must be complete, transitive and context-independent. If a cost-benefit analyst is to use a person's purified preferences as indicators of that person's well-being in arriving at policy recommendations, then (at least within the relevant policy domain) purified preferences must have those same properties. So for Hausman's proposal to work, each individual's undistorted reasoning must generate latent preferences that are complete, transitive and context-independent. But can we expect this to be the case?

In Hausman's analysis of rational preference formation, the agent is represented as engaging in sound reasoning about the truth or falsity of propositions. A preference is a particular kind of proposition – a total subjective comparative evaluation – that the agent

holds to be true. This conceptual framework allows a definition of 'distorted' or 'mistaken' reasoning as reasoning that contravenes principles of conceptual coherence or valid inference, broadly understood. Since preference purification removes the effects of such reasoning, it will result in a set of preference propositions that are consistent with one another and with other propositions to which the agent assents. Since preferences may be derived from subjective propositions (for example, judgements about what is desirable or about the constituents of well-being), this account of preference formation preserves the subjectivity of preferences. As we have explained, Hausman is able to argue that sets of preference propositions that violate transitivity or context-independence are inconsistent and hence incapable of being generated by sound reasoning from consistent premises. But he explicitly denies that sound reasoning necessarily generates a preference relation that is complete (p. 19); all he can say is that completeness is necessary *if choices are to be determined by preferences*.

Thus, Hausman's analysis does not resolve the problem we identified in the literature of behavioural welfare economics. That problem was to justify the implicit assumption that, for any given individual, there exists some mode of latent reasoning that generates complete and context-independent subjective preferences. Were there such a mode of reasoning, it could be argued that context-dependent choices are the result of reasoning imperfections – that is, of failure to recognise the implications of sound reasoning. But Hausman's analysis leaves open the possibility that there are pairs of options for which sound reasoning is unable to determine a preference ranking. The implication is that context-dependent choices are not necessarily mistakes that can be corrected by purification.

4. The inner rational agent

As we noted in Section 1, Hausman is less favourably disposed to nudge policies than are most contributors to behavioural welfare economics. His reservations about nudging, presented in a jointly-authored paper (Hausman and Welch, 2010; henceforth 'HW'), throw light on the model of agency that underlies his analysis of preference purification.

In their opening summary of libertarian paternalism, HW express agreement with many of Sunstein and Thaler's conclusions *about welfare*. For example, in relation to TS's argument for nudging people to save more for retirement, HW say that TS 'are impressed by the imperfections in individual decision-making illustrated by the extent to which people's choices to save for retirement are influenced by details concerning enrolment that ought to be

of negligible importance', and that TS 'catalogue many factors that can lead to mistakes in human judgment and decision-making'. For the purposes of their paper, HW do not need to defend TS's specific judgements about 'which factors interfere with rational deliberation', but they endorse those judgements as 'generally plausible'.

HW's reservations about libertarian paternalism are not about its analysis of welfare, but about the nudging policies that it recommends. These reservations are formulated in terms of *autonomy*, defined as 'the control an individual has over his or her evaluations and choices'. If one is concerned about autonomy, HW say, 'there does seem to be something paternalistic, not merely beneficent, in designing policies so as to take advantage of people's psychological foibles for their own benefit' (p. 128). Throughout the paper, nudges are contrasted with 'rational persuasion'. For example:

> The reason why nudges such as setting defaults seem … to be paternalist, is that in addition to or apart from rational persuasion, they may 'push' individuals to make one choice rather than another… [W]hen this 'pushing' does not take the form of rational persuasion, their autonomy – the extent to which they have control over their own evaluations and deliberations – is diminished. Their actions reflect the tactics of the choice architect rather than exclusively their own evaluation of alternatives. (p. 128)

And (having defined 'shaping' as 'the use of flaws in human decision-making to get individuals to choose one alternative rather than another' [p. 128]):

> [R]ational persuasion respects both individual liberty and the agent's control over her own decision-making, while, in contrast, deception, limiting what choices are available or shaping choices risks circumventing the individual's will. (p. 130)

But what do HW mean when they refer to 'the individual' or 'the agent' as an entity that may or may not have control over his or her evaluations, deliberations and choices? Notice that this agent is not a real human being whose thoughts and actions are governed by psychological mechanisms. If the choices of the real human being are influenced by factors that cannot be construed as good reasons, HW are able to claim that this agent's will has been circumvented. The implication seems to be that the agent is capable of error-free autonomous reasoning that is undistorted by 'problematic' human psychological mechanisms. It is open to rational persuasion, but impervious to attempts to influence it by other means. Given any decision problem, it can identify the option that it wishes to choose, referring 'exclusively' to its own evaluations of alternatives. This seems to imply that the agent's reasoning can generate complete, transitive and context-independent total comparative evaluations. We will call this disembodied entity the *inner rational agent*.

Notice how ordinary human psychology is being treated as a set of forces that are liable to restrict the inner agent's ability to act according to the implications of its own reasoning. It is as if the inner rational agent is separated from the world in which it wants to act by a *psychological shell*. The human being's behaviour is determined by interactions between the autonomous reasoning of the inner agent and the psychological properties of the outer shell. However, in relation to issues of preference and judgement, the inner agent is the ultimate normative authority.

Something like this model of human agency seems to be implicit in Hausman's account of preference purification. Preference purification can be thought of as an attempt to reconstruct the preferences of the inner rational agent by abstracting from the distorting effects of – by 'seeing through' – the psychological shell. We suggest that behavioural welfare economics rests on a similar model of agency, albeit with a less fully-developed account of the reasoning of the inner rational agent. Recall that Sunstein and Thaler's criterion of well-being is given by the preferences that the relevant individual would reveal, were she to pay full attention to decision problems and to possess complete information, unlimited cognitive abilities, and complete self-control. One might think of these preferences of those of an inner rational agent whose reasoning is free of *internal* errors but which depends on faulty psychological machinery to provide it with information, to carry out complex information-processing operations, and to execute its decisions. Lack of attention can cause faults in the flow of information to the inner agent; limited cognitive ability can cause faults in information processing; lack of self-control can cause faults in decision execution. Preference purification is an attempt to reconstruct the decisions that the inner agent would execute if the faults in the psychological shell were corrected.

5. Is the model of the inner rational agent tenable?

Let us begin by recording our surprise that so many behavioural economists have wanted to use the model of the inner rational agent. One of the first impulses for what is now called behavioural economics was a recognition that the mental processes that people actually use in decision-making do not necessarily generate choices with the rationality properties traditionally assumed in economics. An obvious corollary of this idea, pointed out by Kahneman (1996), is that rational choice is not self-explanatory: cases in which behaviour is consistent with the conventional theory of rational choice are just as much in need of psychological explanation as are deviations from that theory. The model of the inner rational

agent seems to depend on a denial of this corollary. In that model, human psychology is represented as a set of forces which affect behaviour by *interfering with* rational choice, but rational choice itself – represented by the error-free reasoning of the inner agent – is not given any psychological explanation. Kahneman is right to say that this modelling strategy is 'deeply problematic' (pp. 251–252).[5]

It might be objected that the model of the inner rational agent has psychological foundations in dual-process theories of the mind. The idea that the workings of the mind can be separated into two 'systems' – the fast and automatic *System 1* and the slow and reflective *System 2* – has been suggested by a number of psychologists (e.g. Wason and Evans, 1975; Kahneman, 2003), and is the central theme of Kahneman's (2011) overview of his contributions to psychology and behavioural economics. Since Sunstein and Thaler use the same idea as an organising principle when reviewing behavioural findings (TS: 19–39), it is plausible to conjecture that they are thinking of the inner rational agent as System 2 and the psychological shell as System 1.

Recently, there has been something of a fashion for economists to appeal to dual-process neurological theories to motivate models of time-inconsistent behaviour. In these models, individual behaviour is determined by interactions between two neural systems – one far-sighted, rational and strategically sophisticated, the other either short-sighted and naïve or automatic. For example, Bernheim and Rangel (2004) explain the decision-making of drug addicts in terms of interactions between a 'cold' mode of reasoning, capable of solving dynamic stochastic programming problems, and a 'hot' mode of automatic responses to environmental cues. Benhabib and Bisin (2005) use a similar model of interaction between 'controlled' and 'automatic' processes to explain consumption and saving behaviour. Fudenberg and Levine (2006) and Brocas and Carillo (2008) present models in which both systems are represented as rational maximisers, but one is far-sighted and the other is myopic. This dual-process modelling strategy is perhaps reasonable as a way of representing the decision-making of the type of drug user who repeatedly tries and fails to quit, or of the former drug user who consciously tries to avoid cues that might induce recidivism. One might be more sceptical when the same strategy is applied to everyday cases of preference

[5] One might also see it as a methodologically questionable attempt to conserve the neoclassical theory of rational choice in the face of disconfirming evidence by re-interpreting it as applying, not to real human beings, but to imaginary disembodied agents. Compare Berg and Gigerenzer's (2010) critique of 'as-if behavioural economics'. See also Goldstein and Gigerenzer's (2002: 75) criticism of work in psychology that treats heuristics as 'poor surrogates for optimal procedures'.

inconsistency, as when Fudenberg and Levine explain the high levels of risk aversion observed in laboratory experiments by hypothesising that the typical student subject has deliberately constrained her own access to cash as a way of solving a self-control problem.

But even if one accepts the dual-process theory as a useful way of organising ideas about human psychology, the model of the inner rational agent remains vulnerable to Kahneman's (1996) critique. One is not entitled simply to assume that the mental processes of System 2 can generate preferences and modes of strategic reasoning that are consistent with conventional decision and game theory. Indeed, that assumption does not fit easily with the logic of dual-process theory. One of the fundamental insights of that theory is that the automatic processing mechanisms of System 1 are evolutionarily older than the conscious mechanisms of System 2. Thus, except in so far as its original features have atrophied, we should expect System 1 to be capable of generating reasonably coherent and successful actions without assistance from other processes. But if System 2 processes are later add-ons, there is no obvious reason to expect them to be able to work independently of the processes to which they have been added. Kahneman (2011: 24) hints at the subsidiary role of System 2 when he says that '[w]hen System 1 runs into difficulty, it calls on System 2 to support more detailed and specific processing that may solve the problem of the moment'. The suggestion seems to be that, in dealing with choice problems, System 2's primary role is to provide decision support services. It is not obvious that this system always needs to be capable of making decisions in its own right, as the inner rational agent is supposed to be.

We have suggested that there can be choice problems that lack determinate rational solutions (with the implication that System 2 may not be able to solve them). To explore this possibility further, we return to the case of SuperReasoner in the cafeteria. Recall that SuperReasoner is a re-engineered version of an ordinary human being, Joe. He differs from Joe by not being subject to reasoning imperfections: he has no limitations of attention, information, cognitive ability or self-control. In all other respects, however, he is the same as Joe. According to Sunstein and Thaler, SuperReasoner's choices reveal Joe's latent preferences. Suppose that the options available at the cafeteria include cream cake and fruit. Were Joe to go to the cafeteria, he would choose (and would be willing to pay a small premium for) whichever of those two options was displayed more prominently. The cafeteria director has read *Nudge*, and wants to use the display that best satisfies Joe's latent preferences. Thus, she needs to know what SuperReasoner would choose. This raises the

question that we asked (but did not answer) in Section 2: Would the probability of his choosing cake be independent of the position of cake on the counter?

One way of answering this question builds on Bacharach's (1993, 2006) analysis of frames. Joe's preferences are context-dependent because the problem of choosing between food items can be framed in different ways. Or, more accurately: Joe can *represent the problem to himself* in different ways. In one case, he construes the problem as a choice between (say) 'the cake at the front of the counter' and 'the apple at the back'; in another, he construes it as a choice between 'the apple at the front' and 'the cake at the back'. In the first case, he prefers 'the cake at the front'; in the second, he prefers 'the apple at the front'. Joe's preferences, *as viewed by Joe himself*, are not inconsistent. They are inconsistent only as viewed by a theorist who conceptualises 'the cake (or apple) at the front' as the same thing as 'the cake (or apple) at the back'. As Bacharach (2006: 13) puts it, whether a decision-theoretic principle of rationality has been violated 'depends on how *we*, the theorists, "cut up the world"'. But, he goes on: 'For decision theory, there are no unproblematically given "same things"'. If this is right, decision theory cannot legitimate the assumption that there is a single correct way of framing the cafeteria problem, accessible to any agent who, like SuperReasoner, is free of reasoning imperfections.

If SuperReasoner's rationality is interpreted in terms of conventional decision theory, as Sunstein and Thaler perhaps intend, Bacharach's argument implies that SuperReasoner's preferences can be context-dependent. However, that argument has less force against Hausman's account of reasoning. Recall that, in that account, sound reasoning can recognise that certain factors 'ought to be irrelevant' for total comparative evaluations. To see where this approach leads, let us stipulate that the relative position of the food items on the counter is such a factor. So SuperReasoner cannot say that, all things considered, the cake is more valuable than the fruit if the cake has the more prominent display, but less valuable in the opposite case. Thus, if we accept Hausman's definition of 'preference', SuperReasoner cannot hold context-dependent *preferences* between fruit and cake. But, since his *feelings* are the same as Joe's, he feels an inclination to choose the cake in the first case, and to choose the fruit in the second. Were his Einstein-like powers of reasoning to lead him to the conclusion that the fruit was more valuable, his Gandhi-like powers of self-control would allow him to overcome any inclination to choose the cake. But let us suppose that, given the premises on which his reasoning operates, the relative value of cake and fruit is undetermined.

If, as we have argued, latent preference is a subjective concept, this supposition does not seem to imply any contradiction. It is true that, by virtue of his special powers, SuperReasoner can access all the information that is relevant for the choice between fruit and cake. For example, he knows all the respects in which eating fruit would improve his health, and all the respects in which eating cake would give him immediate enjoyment. If the uniquely correct choice could be determined by applying some well-defined algorithm to this multi-dimensional information, SuperReasoner would have the computational powers to solve the problem. But we know of no argument, either in behavioural economics or in the theory of rational choice, that would justify the assumption that such an algorithm exists.

So let us maintain our supposition: SuperReasoner cannot determine whether, all things considered, cake is more valuable than fruit or vice versa. In Hausman's sense, he has no (strict or weak) preference between these options. Still, he feels an inclination to choose whichever of cake or fruit is more prominently displayed. What principle of sound reasoning would he contravene by acting on this inclination, just as Joe would? If the answer is 'None', as we believe it is, then SuperReasoner's choices, like Joe's, can be context-dependent.

If instead we are to insist that SuperReasoner's choices must be context-independent, we seem to need to make completeness of preferences an *axiom* of reasoning, rather than a property that, depending on circumstances, reasoning may show to be true or false. Building on Hausman's characterisation of completeness as a boundary condition on rational choice, one might perhaps stipulate that, if an agent is to be truly rational, his choices must always be justified by preferences. One might then claim that rationality requires the agent to ensure that the set of preference propositions he holds to be true is sufficient to pick a nonempty set of justified choices from any nonempty set of options. Our own view (and, apparently, Hausman's) is that this requirement would be unwarranted; but let us set these reservations aside.[6] If SuperReasoner wanted to comply with the requirement, he would have to fill in the gaps in his otherwise incomplete preference ordering by constructing additional preferences whose content was not justified by reasoning.

[6] Some readers may be tempted to think that this requirement could be justified by a 'money pump' (or 'Dutch book') argument, but that thought would be mistaken. Invulnerability to money pumps does *not* imply that choices are determined by preferences (Cubitt and Sugden, 2001). As a simple counter-example, consider an agent who acts on the decision heuristic of never making exchanges, whatever her initial endowment and whatever options are available to her. This heuristic implies a pattern of choice that cannot be rationalised by any (reference-independent) preference relation, but which is clearly invulnerable to money pumps.

But this conclusion is of no help to behavioural welfare economics. The problem that needs to be solved is that of discovering Joe's latent preference between fruit and cake. The line of argument we are exploring leads to the conclusion that, were Joe truly rational, he would have *some* context-independent preference between the two options. But that means only that the imaginary SuperReasoner would have responded to the demands of rationality by constructing such a preference, arbitrarily if necessary. We may have no way of discovering what that imaginary preference would be. And it is only in the most tenuous sense that this imaginary preference is latent *in Joe*.

Savage's (1954: 101–103) discussion of the Allais Paradox nicely illustrates the issues involved here. Savage reports that, when he was first presented with Allais' two choice problems, he expressed a preference for Gamble 1 in Situation 1 and for Gamble 4 in Situation 2 – the response that constitutes the Paradox and that is inconsistent with Savage's own expected-utility axioms (one of which is an axiom of completeness). He confesses that he 'still feel[s] an intuitive attraction to those preferences'. However, since his analysis of expected utility is intended as a normative theory, it would be an 'intolerable discrepancy' for him to maintain two preferences that together were inconsistent with the axioms of the theory:

> 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. (p. 102)

Savage reassures himself of the validity of his axioms by re-framing the four gambles so that their outcomes all depend on the same draw from a set of lottery tickets numbered 1–100. Prizes are assigned to tickets so that the prizes for Gambles 1, 2, 3 and 4 respectively, in units of \$100,000, are (5, 0, 5, 0) for ticket 1, (5, 25, 5, 25) for tickets 2–11, and (5, 5, 0, 0) for tickets 12–100.[7] Since, in each situation, the two gambles on offer differ only in the event that one of tickets 1–11 is drawn, Savage concludes that the other tickets are irrelevant to the decisions that have to be made. Conditional on this event, Gambles 1 and 3 are identical, as are Gambles 2 and 4. Thus, Savage's original preferences are unacceptably context-dependent. Both of them cannot be right. But which of them is wrong? Savage tells himself that, in both situations, the choice problem reduces to 'whether I would sell an outright gift of \$500,000 for a 10-to-1 chance to win \$2,500,000'. Consulting his 'purely personal taste', he

[7] The implicit claim that the original and revised versions of the problems are equivalent to one another is open to question, but it is an implication of Savage's axioms.

finds that he prefers the former. He then accepts the implication that he prefers Gamble 3 to Gamble 4, saying: 'It seems to me that in reversing my preferences between Gambles 3 and 4 I have corrected an error'.

Notice that Savage has invoked a third situation (let us call it 'Situation 3') in which he has to choose *either* \$500,000 with probability 1 ('Gamble 5') *or* \$2,500,000 with probability 10/11 and nothing with probability 1/11 ('Gamble 6'). According to his axioms, his ranking of Gamble 3 relative to Gamble 4 (and, equivalently, his ranking of Gamble 1 relative to Gamble 2), should be the same as his ranking of Gamble 5 relative to Gamble 6. He finds an inclination to prefer Gamble 5 to Gamble 6. So far, this is not a resolution of the original problem; it is merely an expansion of the set of inconsistent preferences. However, it seems that Savage feels more confident about his inclinations in Situation 3 than in the other two situations, and so decides to use those inclinations as his arbiter. There is nothing wrong with that: as Savage says, this is a matter of reflection, not deduction. But there seems no reason to suppose that this particular sequence of reflections leads to the uniquely correct resolution of the original inconsistency (if inconsistency it is). At most, this story tells us that if someone genuinely accepted the expected-utility axioms as requirements of rationality and was not cognitively constrained, he would be able to settle on *some* preferences, consistent with those axioms, which he was willing to live with (but which might still be contrary to his actual inclinations). That is not particularly helpful if we are trying to identify the actual latent preferences of an ordinary Joe whose choices and inclinations have the Allais Paradox pattern.

6. Purification – or regularisation?

We have argued that latent preference is not a useful concept for normative economics. How then (a sceptical reader might ask) can we explain the fact that so many behavioural economists have wanted to use it? We suggest that this practice may be a by-product of a modelling strategy that is common in behavioural economics. When used in the development of descriptive theories, this strategy has significant methodological virtues, but it is liable to lead one astray in normative work.

This strategy, which we will call *behavioural optimisation*, uses conventional rational-choice theory as a template, and models the individual as maximising a *behavioural utility* function that retains many of the properties of the utility functions used in neoclassical economics and game theory. Psychological factors that are neglected in conventional theory

are modelled by allowing behavioural utility to depend on additional variables. Often, the standard utility function is represented as a special case of the behavioural function. Failing that, the two functions can usually be thought of as distinct special cases of a more general utility function.

If the objective is to develop a parsimonious descriptive theory that generates successful predictions about economic behaviour, this strategy has obvious practical merits. If one accepts (as many behavioural economists do) that the predictions of conventional economic theories are often good first approximations to the truth, it may be more productive to look for incremental improvements to those theories than to start again from scratch and to re-invent wheels. Even if one is sceptical about the predictive success of conventional theory, it remains true that economists have developed a large body of abstract theoretical results that hold for maximising behaviour in general, and which can be re-used in behavioural utility models. Using behavioural utility functions also makes it easier to identify and test the novel implications of behavioural theories, and to measure the increase in explanatory power that can be attributed to the inclusion of additional variables. Exactly these arguments are used by Rabin (2013) to defend the behavioural optimisation strategy. Similarly, Hausman (2012: 114–115) favours the strategy of modelling psychological factors such as framing through their effects on preferences on the grounds that 'if economists and decision theorists continue to regard preferences as determinative [of choices], then they can still employ consequentialist and game-theoretic models and the mathematical tools that permit predictions to be derived from them'.[8]

Notice that for the behavioural optimisation strategy to have these merits, it is not necessary that the standard theory is a representation of *rational* choice; what matters is that it makes at least moderately accurate predictions across a wide domain. However, because the standard theory is usually interpreted in terms of rationality, it is tempting to think that this modelling strategy allows one to isolate the effects of mistakes (i.e. those effects on utility that occur because 'behavioural' variables do not take the values that correspond with the standard theory), and so to identify latent preferences (i.e. the preferences that would result if behavioural variables took their standard values). Rabin (2013: 529) presents this

[8] In saying this, Hausman is in danger of undercutting his earlier argument that, in the practice of economics, preferences are implicitly interpreted as total subjective comparative evaluations (see Section 3 above). If there are pragmatic reasons for behavioural economists to use preference-based models when explaining non-rational choices, neoclassical economists might favour preference-based models for pragmatic reasons too.

feature of behavioural utility models as an important merit, on the grounds that it allows us to 'capture many errors in terms of systematic mistakes in the proximate value function … or where the beliefs imported into their maximizations are systematically distorted'. The use of behavioural optimisation models to purify preferences was discussed in Section 2, where it was exemplified by the work of Bleichrodt et al. ('BPW', 2001), Kőszegi and Rabin (2007, 2008), and Bordalo et al. (2013).

We will argue that this method of defining and identifying latent preferences is unsatisfactory. We will develop this argument by considering how BPW's purification methodology might be applied to Savage's version of the Allais Paradox.

Recall that BPW use cumulative prospect theory as the descriptive model of choice. Viewed in the perspective of that theory, Allais' four gambles can be differentiated in terms of two characteristics – the probability of winning at least $500,000, and the probability of winning $2,500,000. In terms of the second characteristic, Situations 1 and 2 are equivalent to one another. (In each situation, the probability of winning $2,500,000 is either zero or 0.10, depending on whether the first or the second gamble is chosen.) So an explanation of the Allais Paradox must work through the first characteristic. The probability of winning at least $500,000 is 1.00 in Gamble 1, 0.99 in Gamble 2, 0.11 in Gamble 3, and 0.10 in Gamble 4. The difference between the two relevant probabilities (1.00 and 0.99 in Situation 1, 0.11 and 0.10 in Situation 2) is the same in both situations, which is another way of explaining why the Paradox contravenes expected utility theory. However, cumulative prospect theory transforms each objective probability p into a subjective decision weight $w(p)$. The Allais Paradox is possible if $w(1.00) - w(0.99)$ is sufficiently greater than $w(0.11) - w(0.10)$. That inequality is consistent with most empirical estimates of the probability weighting function, and also with intuition: the difference between the certainty of a very large prize and a 99 per cent chance of it *feels* more significant than the difference between an 11 per cent chance and a 10 per cent chance. So it is plausible to suppose that cumulative prospect theory is picking up a psychological mechanism that contributes *in some way* to the Allais Paradox.

BPW's purification methodology treats the non-linearity of the probability weighting function as a reasoning error that needs to be corrected if we are to identify latent preferences. But where is the error? Of course, there would have been an error *if* the decision-maker had known the utility to him of the three possible outcomes, *and if*, believing expected utility theory to be the right normative model, he had tried to calculate the expected utility of each of the four gambles, *and if* in doing so he had used decision weights in the mistaken belief

that they were objective probabilities. But that is not a remotely plausible account of the reasoning that leads real people to choose Gambles 1 and 4. To point to just one problem with this account, remember that when people respond to Allais' problems, they are *told* all the relevant objective probabilities. If you were to ask a respondent what he believed to be the percentage probability of an outcome that he had just been told had a probability of 1 per cent, what answer would you expect to get?

What BRW's purification methodology reveals is that, *relative to the benchmark of expected utility theory*, the person who has made the Allais Paradox choices has behaved *as if* he held false beliefs about the probabilities of the relevant events. If expected utility theory could be interpreted as a first approximation to a true description of how people actually reason, it might be plausible to move from that as-if proposition to the conjecture that the person's actual reasoning followed the logic of expected-utility reasoning but with false beliefs. But the truth is surely that expected utility theory provides a first approximation to *the choices that people actually make*, not to the reasoning by which they arrive at those choices.

It is not surprising that expected utility theory has this approximation property, at least when applied to lotteries with monetary outcomes and explicit objective probabilities. Whatever mental processes people use in decision-making about such lotteries, one would expect larger money prizes to be perceived more favourably than smaller ones, other things being equal. Similarly, for any given money amount x, one would expect larger probabilities of winning at least x to be perceived more favourably than smaller probabilities. By generalising these two intuitions and by organising them in a simple and tractable functional form, expected utility theory picks up some of the main patterns in the decisions that are generated by actual human reasoning. In the case of the Allais Paradox, however, cumulative prospect theory provides a more accurate description of actual decisions. In the absence of a theory of how people reason, that is just about all that can be said. One is certainly not entitled to infer that Allais Paradox choices reveal errors of reasoning that are not committed by people whose choices are consistent with expected utility theory.

We conclude that BPW's methodology does not reconstruct latent preferences. In fairness, however, it should be acknowledged that BPW sometimes justify this methodology in more pragmatic terms, as when they say:

> We are well aware that many of the assumptions underlying our proposal are controversial, such as the very existence of true underlying preferences. These

> assumptions are, however, the best that we can think of in the current state of the art for situations where decisions have to be taken, as good as possible, on the basis of quick and dirty data. (p. 1500)

Recall that BPW's paradigm decision problem is that of a professional specialist who has to make a choice on behalf of a client, given only partial information about the client's revealed or stated preferences. When BPW say that expected utility theory is 'the right normative model for choice under uncertainty' (pp. 1498–1499), they seem to be referring to the decision problem faced by the professional. One might perhaps argue that, if the professional shares BPW's view of the normative status of expected utility theory, she ought to construct *her* judgements about the client's welfare, and hence about the decisions *she* should make when acting on the client's behalf, so that these judgements are consistent with the expected-utility axioms. Viewed in this way, what seems to be required is not an inference about the hypothetical choices of the client's inner rational agent, but rather a way of *regularising* the available data about the client's preferences so that it is compatible with the particular model of decision-making that the professional wants to use.

Regularisation in this sense is almost always needed when a theoretical model comes into contact with real data. For example, consider an economic model of the spatial distribution of unemployment. Suppose that, in this model, every job-seeker and every job offer has a spatial location. In the world of the model, this makes perfectly good sense: each job-seeker has a 'home' and each job has a 'workplace'. But if we try to apply the model in practice, we will find that 'home' and 'workplace' can be ambiguous concepts. Some people have two or more home addresses, while some have none; and analogously for jobs. In order to regularise the data so that they fit the model's categorisation scheme, some more or less arbitrary classifications will need to be made. But one would surely not claim that these classifications correspond with latent truths about the world that real job-seekers and real employers have failed to recognise. In the same way, a medical decision-maker might reasonably use BPW's methodology to construct a tractable *model* of the client's preferences, regularised so as to be consistent with expected utility theory, without claiming that the preferences in the model were latent in the client. The arguments we have developed in this paper would not be objections to a version of behavioural welfare economics that claimed only to regularise revealed preferences that were inconsistent with conventional theory, without interpreting this process as the identification and correction of errors, or as a way of helping individuals to make better choices. But that is *not* the version of behavioural welfare economics that is to be found in the literature.

7. Conclusion

In arguing for libertarian paternalism, Sunstein and Thaler (TS: 6) criticise conventional economists for assuming that ordinary human beings are 'Econs' – an imaginary species which 'thinks and chooses unfailingly well'. Sunstein and Thaler claim that their own approach to behavioural welfare economics – an approach that is becoming part of the mainstream of behavioural economics, and whose core features are endorsed by Hausman (2012) – breaks away from this mistaken assumption, and models human psychology as it really is. We have argued that this claim is misleading. It would be closer to the truth to say that behavioural welfare economics models human beings as faulty Econs. Its implicit model of human decision-making is that of a neoclassically rational inner agent, trapped inside and constrained by an outer psychological shell. Normative analysis is understood as an attempt to reconstruct and respect the preferences of the imagined inner Econ.

We maintain that if behavioural and normative economics are to be satisfactorily reconciled, the first essential is that economists learn to live with the facts of human psychology. We need a normative economics that does not presuppose a kind of rational human agency for which there is no known psychological foundation. One possible line of advance is to find a normative criterion that respects individuals' choices without referring to the preferences – consistent or inconsistent – that lie behind them. Sugden's (2004) 'opportunity criterion' is an example of this strategy. Such a criterion may seem unappealing if one presupposes that normative economics is addressed to a benevolent social planner, but this addressee is no more than a theoretical or literary construct. If one thinks of individual citizens as principals and public decision-makers as their agents, it may seem more natural to treat *citizens* as the addressees of normative economics. Citizens who recognise that their choices are sometimes context-dependent might still want their agents to respect those choices (Sugden, 2013).

To readers who would prefer to conserve more of the framework of conventional welfare economics, we commend the 'regularisation' perspective that we sketched in Section 6. Instead of claiming to reconstruct the latent neoclassical preferences of individuals whose psychology has led them into error, economists might think of themselves as doing their best to represent the complex reality of human judgement and decision-making in a highly simplified but perhaps still useful normative modelling framework. But that would require a significant retreat from the ambition – not to say hubris – of much of the current literature of behavioural welfare economics.

References

Bacharach, Michael (1993). Variable universe games. In Ken Binmore, Alan Kirman, and Piero Tani (eds) Frontiers *of Game Theory*. Cambridge, MA: MIT Press. pp. 254–275.

Bacharach, Michael (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Edited by Natalie Gold and Robert Sugden. Princeton, NJ: Princeton University Press.

Benhabib, Jess and Alberto Bisin (2005). Modeling internal commitment mechanisms and self-control: a neuroeconomics approach to consumption–saving decisions. *Games and Economic Behavior* 52: 460–492.

Berg, Nathan and Gerd Gigerenzer (2010). As-if behavioral economics: neoclassical economics in disguise? *History of Economic Ideas* 18: 133–166.

Bernheim, Douglas and Antonio Rangel (2004). Addiction and cue-triggered decision processes. *American Economic Review* 94: 1558–1590.

Bernheim, Douglas and Antonio Rangel (2007). Toward choice-theoretic foundations for behavioral welfare economics. *American Economic Review: Papers and Proceedings* 97: 464–470.

Bernheim, Douglas and Antonio Rangel (2009). Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics* 124: 51–104.

Bleichrodt, Han, Jose-Luis Pinto-Prades and Peter Wakker (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science* 47: 1498–1514.

Bordalo, Pedro, Nicola Gennaioli and Andrei Shleifer (2013). Salience and consumer choice. *Journal of Political Economy* 121: 803–843.

Brocas, Isabelle and Juan Carrillo (2008). The brain as a hierarchical organization. *American Economic Review* 98: 1312–1346.

Broome, John (1991). Utility. *Economics and Philosophy* 7: 1–12.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donaghue and Matthew Rabin (2003). Regulation for conservatives: behavioral economics and the case for 'asymmetric paternalism'. *University of Pennsylvania Law Review* 151: 1211-1254.

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

Fudenberg, Drew and David Levine (2006). A dual-self model of impulse control. *American Economic Review* 96: 1449–1476.

Goldstein, Daniel and Gerd Gigerenzer (2002). Models of ecological rationality: the recognition heuristic. *Psychological Review* 109: 75–90.

Halpern, David and Max Nesterak (2014). Nudging the UK: a conversation with David Halpern. *The Psych Report* (5 January). http://thepsychreport.com/conversations/nudging-the-uk-a-conversation-with-david-halpern/

Hansen, Pelle G. and Andreas M. Jespersen (2013). Nudge and the manipulation of choice: a framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation* 2013(1): 3–28.

Hausman, Daniel (2012). *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.

Hausman, Daniel and Brynn Welch (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy* 18: 123–136.

Kahneman, Daniel (1996). Comment [on Plott (1996)]. In Kenneth Arrow, Enrico Colombatto, Mark Perlman and Christian Schmidt (eds), *The Rational Foundations of Economic Behaviour* (Basingstoke: Macmillan and International Economic Association), pp. 251–254.

Kahneman, Daniel. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*. 58: 697–720.

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

Kahneman, Daniel, Jack L. Knetsch and Richard H. Thaler (1990). Experimental tests of the endowment effect and the Coase Theorem. *Journal of Political Economy* 98: 1325–1348.

Kőszegi, Botond and Matthew Rabin (2007). Mistakes in choice-based welfare analysis. *American Economic Review* 97: 477–481.

Kőszegi, Botond and Matthew Rabin (2008). Choices, situations, and happiness. *Journal of Public Economics* 92: 1821–1832.

Li, Chen, Zhihua Li and Peter Wakker (2014). If nudge cannot be applied: a litmus test of the readers' stance on paternalism. *Theory and Decision* 76: 297–315.

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

Rabin, Matthew (2013). Incorporating limited rationality into economics. *Journal of Economic Literature* 51: 528–543.

Read, Daniel and Barbara van Leeuwen (1998). Predicting hunger: the effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes* 76: 189–205.

Salant, Yuval and Ariel Rubinstein (2008). (*A*, *f*): choice with frames. *Review of Economic Studies* 75: 1287–1296.

Savage, Leonard (1954). *The Foundations of Statistics*. New York [?]: Wiley.

Schkade, David and Daniel Kahneman (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science* 9: 340-346.

Sunstein, Cass (2014). The ethics of nudging. Available at SSRN: http://ssrn.com/abstract=2526341

Sunstein, Cass R. and Richard Thaler (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review*, 70: 1159-1202.

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

Sugden, Robert (2013). 'The behavioural economist and the social planner: to whom should behavioural welfare economics be addressed?' *Inquiry* 56: 519–538.

Thaler, Richard and Cass Sunstein (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

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

Wason, Peter and Jonathan Evans (1975). Dual processes in reasoning? *Cognition, 3*, 141–154.

Wilson, Chris and Catherine Waddams Price (2010). Do consumers switch to the best supplier? *Oxford Economic Papers* 62: 647–668.