Looking for a psychology for the inner rational agent

Robert Sugden

Centre for Behavioural and Experimental Social Science and School of Economics
University of East Anglia
Norwich NR4 7TJ
United Kingdom
r.sugden@uea.ac.uk

1 June 2015

Abstract: Research in psychology and behavioural economics shows that individuals' choices often depend on 'irrelevant' contextual factors. This presents problems for normative economics, which has traditionally used preference-satisfaction as its criterion. A common response is to claim that individuals have context-independent latent preferences which are 'distorted' by psychological factors, and that latent preferences should be respected. This response implicitly uses a model of human action in which each human being has an 'inner rational agent'. I argue that this model is psychologically ungrounded. Although references to latent preferences appear in psychologically-based explanations of context-dependent choice, latent preferences serve no explanatory purpose.

Keywords: inner rational agent; behavioural welfare economics; preference purification; attention; true self

Acknowledgements: A previous version of this paper was presented at a workshop on the topic 'Respecting context-dependent preferences' at Umeå University in 2014. I thank participants at that workshop, Sudeep Bhatia, Kalle Grill, Graham Loomes, Danny Scoccia and an anonymous referee for valuable comments. My work was supported by the Economic and Social Research Council through the Network for Integrated Behavioural Science (grant reference ES/K002201/1).

A recurring finding of behavioural economics is that individuals' choices between what might naturally be thought of as given outcomes can vary according to apparently irrelevant features of the context in which those choices are made. For example, faced with a choice between a specific amount of money and a specific consumer good, people are less likely to choose the money if the decision is framed in terms of selling something that they own than if it is framed as a straight choice (Kahneman, Knetsch and Thaler, 1990). When choosing between alternative snacks to be delivered at a fixed time a week in the future, people are more likely to choose unhealthy but hunger-satisfying items if they are hungrier at the time they make the decision (Read and van Leeuwen, 1998). In calling such contextual features 'irrelevant', I mean that they have no obvious relevance to the decision-maker's well-being, interests or goals; changes in these features seem therefore not to provide good *reasons* for a person to change her preferences. Nevertheless, there are well-grounded psychological *explanations* of why revealed preferences are context-dependent. These findings present a problem for normative economics, because there is a long tradition in economics of using preference-satisfaction as the criterion for evaluating alternative policy options. That public decision-makers should respect individuals' preferences has long been an important idea in liberal political philosophy. But should we – indeed, can we – respect context-dependent preferences?

Many economists and philosophers find the idea of respecting context-dependent preferences problematic, either because there seems to be no good reason for thinking that such preferences are indicators of individual well-being, or more fundamentally, because the concept of respecting a person's preferences is thought to be ill-defined unless those preferences satisfy minimal properties of internal consistency.[1] However, the same writers are often reluctant to conclude that there is no need to respect preferences at all, and that public decision-makers should simply use their own best judgements about the effects of policies on individuals' well-being – a conclusion that seems unacceptably paternalistic. A common escape route from this impasse is to argue that individuals whose *choices* are context-dependent are not revealing the *preferences* that in some meaningful sense they

[1] I have argued for an approach to normative economics which attaches value to individuals' opportunities rather than to the satisfaction of their preferences. This approach does not depend on any assumptions about the coherence of individuals' preference while, in a certain sense, respecting whatever preferences individuals act on (Sugden, 2004, 2007; McQuillin and Sugden, 2012). It is therefore not vulnerable to the problems that are the topic of the current paper. However, this approach has not yet found much favour in economics or philosophy.

actually hold, and that their ‘true’, ‘underlying’ or ‘latent’ preferences are context-independent. The disparity between latent preference and choice is attributed to psychological mechanisms which induce systematic biases or errors in reasoning. If these latent preferences could be recovered, it would be possible to use the traditional methods of welfare economics to work out how best to satisfy them. Following Hausman (2012, p. 102), I will call the process of recovering latent preferences *preference purification*. I will call the broader strategy of using such preferences in welfare economics *behavioural welfare economics*. By using this strategy, it is thought, the principle of respect for individuals’ preferences can be retained.

In another paper, Infante, Lecouteux and I (2015) have examined how this strategy has been used by behavioural economists. We present a critique of behavioural welfare economics from the perspective of philosophy of mind. We argue that the strategy understands human agency as if each individual human being has a ‘rational true self’ or ‘inner rational agent’ which has access to some mode of valid reasoning that can generate context-independent preferences. Psychological explanations of context-dependent preferences are then interpreted as if the individual’s psychology was an external force subverting the will of the true self. The inner rational agent is not endowed with any psychology of its own, and no description is given of the mode of reasoning it is supposed to use. We argue that this model of agency is ungrounded and implausible. In Section 1 of the current paper, I summarise that argument.

However, my main aim in the present paper is to consider behavioural welfare economics from a different perspective, that of cognitive psychology. One reason for suspicion about the model of the inner rational agent is that its capacity for correct reasoning is not given any psychological explanation. So one way of trying to make sense of the model is to understand decision-making, both rational and irrational, in terms of psychological mechanisms of mental processing, and to try to isolate some component or aspect of this mental processing that corresponds with rational deliberation and that is capable of generating context-independent preferences. If such a component could be isolated, and if actual behaviour could be represented as the result of interaction between it and other psychological mechanisms, the isolated component might be interpreted as the psychological substrate of the inner rational agent and the other mechanisms as potential causes of error.

In Section 2, I consider this isolation strategy in general terms, and argue that it is unlikely to succeed. A psychological explanation of context-dependent choices does not need a concept of ‘true’ preference. In the most credible of such explanations, responses to contextual cues are an integral part of the mental processes of decision-making. The idea of recovering latent preferences by removing the influence of these cues seems incoherent.

In Sections 3 and 4, I support this general claim by examining two specific models – one from behavioural economics, the other from cognitive psychology – which at first sight might seem to provide clues about how latent preferences can be isolated. Both models represent the role of attention in the mental processes that underlie decision-making. The first model is typical of much current work in behavioural economics in its use of concepts of correct reasoning and latent preferences in relation to what are supposed to be models of mental processing, understood empirically. The second model has a much richer representation of mental processes and is presented with much less – but, interestingly, still with some – reference to correctness of reasoning. I will argue that, in both models, concepts of correctness play no explanatory role. Thus, however successful these models may be in explaining decision-making, they do not provide any empirical grounding for the concept of latent preferences.

1. The model of the inner rational agent[2]

Preference purification is at the core of ‘behavioural welfare economics’ – a method of normative analysis that has been used by many prominent behavioural economists. Infante, Lecouteux and I document the use or advocacy of this method by, among others, Bleichrodt, Pinto-Prades and Wakker (2001), Camerer et al. (2003), Sunstein and Thaler (2003; henceforth ‘ST’), Kőszegi and Rabin (2007), Salant and Rubinstein (2008), Thaler and Sunstein (2008; henceforth ‘TS’), and Bernheim and Rangel (2009). Taking a more philosophical perspective, Hausman (2012, pp. 100–102) gives a qualified endorsement to preference purification as a means of making judgements about individual well-being. In the present paper, I will focus on the particularly influential work of Sunstein and Thaler.

Sunstein and Thaler claim that the findings of behavioural economics make paternalism unavoidable. This claim is developed in relation to the now-familiar example of a cafeteria director choosing how to display food items when she knows that her customers’ choices are influenced by the prominence with which different items are displayed. Characterising their anti-paternalist opponents as advocating that the director should ‘give

[2] This section is based on Infante, Lecouteux and Sugden (2015a, 2015b).

consumers what she thinks they would choose on their own', Sunstein and Thaler claim that the anti-paternalist position is 'incoherent', because the customers lack 'well-formed' (that is, context-independent) preferences. In their 2004 paper, Sunstein and Thaler conclude that the only reasonable decision criterion for the cafeteria director is to 'make the choices that she thinks would make the customers best off, all things considered' (ST, pp. 1164–1165, 1182). In their 2008 book, they make a significant revision to this criterion, declaring that their recommendations are designed to 'make choosers better off, *as judged by themselves*' (TS, p. 5; italics in original). The italicised clause recurs with minor variations throughout TS (e.g. pp. 10, 12, 80). The implication is that the addressee of Sunstein and Thaler's work – originally called the 'planner', but restyled in 2008 as the 'choice architect' – tries to respect each individual's subjective judgements about what makes him better off.

But how are these judgements to be defined, and how can they reconstructed? Sunstein and Thaler are coy about this, but they provide some clues about their thinking when, immediately after presenting the principle of trying to make choosers 'better off, *as judged by themselves*', they 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, p. 5)

The implication is that what makes an individual better off 'as judged by himself' is defined by the preferences he would have revealed, had his decision-making not been affected by limitations of attention, information, cognitive ability or self-control. So Sunstein and Thaler's approach to normative economics treats context-dependent choices as the result of errors of reasoning. It requires the reconstruction of individuals' *latent preferences* by simulating what they would have chosen, had their reasoning not been subject to these errors. This is preference purification. Clearly, this approach can overcome the problem of context-dependence in actual choices only if, as Sunstein and Thaler implicitly assume is the case, the corresponding latent preferences are context-independent.

Behavioural welfare economics can be characterised more precisely as having the following four properties. (1) Behavioural welfare economics is intended to apply to cases in which individuals' revealed preferences depend on contextual factors that have little or no apparent relevance to those individuals' interests or well-being. (2) The normative criterion is the satisfaction of each individual's latent preferences, defined as the preferences he would reveal in the absence of any errors that might be caused by limitations of attention, information, cognitive ability or self-control. (3) Latent preferences are interpreted as expressing individuals' subjective judgements about their interests or well-being; they do not necessarily track objective properties of the external world (such as an individual's monetary wealth or health status) or properties of passive experience (such as happiness in the hedonic sense). (4) In the cases to which behavioural welfare economics is to be applied, latent preferences are assumed to be context-independent.

Infante, Lecouteux and I argue that this approach implicitly uses a model of an *inner rational agent*. By this, we mean that it treats human agency *as if* each human being was made up of a neoclassically rational entity encased in, and able to interact with the world only through, an error-prone psychological shell. Of course, we do not claim that behavioural economists think that human beings are *really* made up of these components. The idea of the inner rational agent is merely a way of making vivid an implication of properties (1) to (4). To be more specific, these properties imply that the human individual has a latent capacity, constant across decision environments, to form context-independent subjective judgements on the basis of error-free reasoning. This capacity is not always revealed in the individual's actual decision-making behaviour, but its effects can be isolated by identifying what the individual would have chosen in the absence of errors. 'The inner rational agent' is our name for that capacity.

In arguing that this model is problematic, Infante, Lecouteux and I direct most of our criticism at the assumption that latent preferences, as defined by the preference purification method, are complete and context-independent. By treating context-dependent choices as revealing errors of reasoning, the preference purification approach implicitly assumes that for each individual there is some mode of latent reasoning which, if carried out correctly, would generate complete and context-independent preferences. If one interprets preferences as subjective propositions that the relevant individual holds to be true, decision theory imposes consistency restrictions on the set of preference propositions that an individual simultaneously holds to be true, but it provides no explanation of how she arrives at those propositions.[3] The advocates of preference purification invoke a concept of 'correct' or 'undistorted' reasoning, latent in the real individual, which can somehow create complete and

[3] Decision theory and game theory are not theories *of reasoning* (that is, theories about the processes by which new propositions are inferred from existing ones); their concepts of 'rationality' are consistency conditions that restrict the sets of propositions that an unmodelled process of reasoning is allowed to generate. For more on this, see Broome, 2013 and Cubitt and Sugden, 2014).

context-independent preferences; but they provide no account of what this reasoning actually is. We accept that a defensible account of correct reasoning might show that *the set of preference propositions that can be derived by that reasoning* must satisfy certain consistency conditions, perhaps including some condition of independence of 'irrelevant' contextual features. But we challenge the implicit assumption that, for every pair of choice objects *x* and *y*, correct reasoning can lead to one of the conclusions '*x* is preferable to *y*', '*y* is preferable to *x*', or '*x* and *y* are equally preferable'. If one accepts that the preference relation that an individual can be derive by correct reasoning may be incomplete, one cannot infer errors of reasoning from context-dependent *choices*. Thus, contrary to a crucial implicit assumption of the preference purification approach, context-dependency in actual choice may recur in the hypothetical choices that are supposed to reveal latent preferences.

2. Trying to make psychological sense of latent preferences

Given that behavioural economists usually characterise their sub-discipline as economics with psychological foundations, it is surprising how little work has been done to explain latent preferences in psychological terms. In the remainder of the paper, I consider whether the concept of latent preference might be given empirical content by interpreting it as a component of a psychological model of mental processing.

Some idea of the difficulties involved in this task can be had by considering an example I mentioned in the Introduction – people's choices between alternative snacks to be delivered a week after that choice was made. What has been found is that a typical individual's choices between specific food items (for example, Mars bars and apples) are influenced by his *current* degree of hunger, even though the date and time of delivery of the snack (and hence, the presumably predictable degree of hunger *at the time of delivery*) is held constant. This is a paradigm case in which choice is influenced by a contextual cue which seems to have no relevance for the individual's welfare. In broad-brush terms, the psychological mechanism behind this effect is easy to understand. Mars bars and apples are goods with different mixes of attributes: the Mars bar is more energy-giving and perhaps (as viewed by the individual) tastier, the apple is more refreshing and (as an addition to the individual's typical diet) healthier. In deliberating about which of the two snacks to choose, the individual has to bring these various attributes to mind and strike a balance between them. The hungrier he is, the more attention he gives to those attributes on which the Mars bar is superior, and so the more likely it is that his deliberation will end in the choice of that option.

Viewed in this way, what might seem to be irrational context-dependence is evidence about the underlying structure of the decision-making mechanism. If one thinks in terms of the evolutionary origins of human psychology, the role played by attention in decision-making can be understood as an integral part of a general-purpose mechanism for choosing between multi-attribute options – a mechanism that is (as if) efficiently designed to make use of other mental processes that tend to distribute attention towards what is currently important. (For example, the hungrier one is, the more important it is to be alert to possible sources of nutrition.)

But how, then, are we to separate the decision-making mechanism into components of 'rationality' and 'error', and to be able to claim that the rational component retains the subjectivity of the real human individual? The only possible way forward that I can see is to try to identify some particular distribution of attention as 'correct'. But how are we to do this? Recall that it is fundamental to the preference purification approach that the individual's latent preferences represent his own subjective judgements. Thus, we cannot define the correct distribution of attention in terms of some *objective* standard of the individual's interest, analogous with fitness in an evolutionary model. In the absence of such a standard, the idea of a 'neutral' distribution of attention between different attributes of choice options is ill-defined. (To mention just one problem, suppose we define neutrality as equal attention to every attribute. In the case of the snacks, is the effect of diet on weight a single attribute, or are health and slimness two separate attributes?) It would be circular to define the correct distribution of attention as that which would generate 'true' latent preferences, since latent preferences have already been defined in terms of correct reasoning.

The core of the problem is that the attention-based mechanisms that explain the individual's decisions also explain what, given the relevant choice context, he actually prefers or desires to do: he feels the desires that prompt him to choose as he does. Viewed in the perspective of empirical psychology, the idea that he might have 'true' preferences that are different from the actual ones seems free-floating and redundant.

So far, I have been arguing in very general terms. I will now try to give further support for my sceptical conclusions by looking at two concrete examples of the use of the concept of latent preference in behavioural economics and psychology.

3. A behavioural economic model of attention

My first case study is chosen as a characteristic example of how the concept of latent preference is used in behavioural economic models. It is the analysis in a recent paper entitled ‘Salience and consumer choice’, by Bordalo, Gannaioli and Shleifer (hereafter ‘BGS’; 2013) and published in the prestigious *Journal of Political Economy*. BGS develop a model that is motivated by experimental findings from psychology, economics and marketing. The core idea is expressed in a quotation from a psychological paper by Taylor and Thompson (1982, p. 175): ‘salience refers to the phenomenon that when one’s attention is differentially directed to one portion of the environment rather than to others, the information contained in that portion will receive disproportionate weighting in subsequent judgements’.

BGS’s core model is of a decision problem in which a consumer faces a choice set containing two or more *goods*, one and only one of which is to be chosen. Each good *k* is characterised by the pair $\langle q_k, p_k\rangle$, where q_k and p_k are non-negative magnitudes, respectively representing the *quality* and *price* of that good. The consumer knows the value of q_k and p_k for each good in the choice set. Higher-quality goods are assumed to have higher prices. In effect, BGS assume that quality is measured in its own units on a ratio scale (i.e. a scale on which the zero point is fixed but the unit of measurement is arbitrary).[4] In BGS’s leading example, the reader is asked to imagine choosing between a bottle of French wine priced at $20 and a bottle of Austrian wine priced at $10 when the reader thinks the French wine ‘is perhaps 50 percent better’ (pp. 803–804). I take it that, in this example, the consumer is thinking of units of quality as categorically different from units of price, and is trying to decide how to make trade-offs between the two attributes. (For example, he might be thinking of the quality scale in terms of the answer he would give to a question asking him to rate the quality of the wine on a scale from 0 to 10; he rates the Austrian wine as 6 and the French wine as 9.)

The crucial assumption of the model is stated as follows:

> Without salience distortions, a consumer values good *k* with a linear utility function, $u_k = q_k - p_k$, which attaches equal weights to quality and price. A salient thinker departs from [this utility function] by inflating the relative weights attached to the attributes that he perceives to be more salient. … [W]e say that an attribute

[4] BGS actually say: ‘Quality and price are measured in dollars and known to the consumer’ (p. 807). Despite the literal meaning of this sentence, I think my interpretation is faithful to BGS’s intentions. It is only because quality and price are measured in different units that the consumer faces a non-trivial choice problem.

> (quality or price) is salient for good *k* in the choice set … if this attribute ‘stands out’ relative to the good’s other attributes. (p. 807)

I take the first sentence to mean that BGS are assuming a weighted linear utility function $u_k = \alpha_Q\, q_k - \alpha_P\, p_k$, and are defining the unit in which quality is measured so that $\alpha_P = \alpha_Q = 1$. This utility function represents the consumer’s latent preference ordering over ⟨quality, price⟩ pairs; these preferences can be described by a family of what BGS call ‘rational indifference curves’, which are linear and parallel. That the marginal rate of substitution between units of price and quality is constant is a substantive modelling assumption; that each unit of quality is worth $1 to the ‘rational consumer’ is merely a convenient normalisation. Unless the consumer is known to be ‘rational’, these indifference curves are not directly revealed in choices. Notice that so far, the concepts of rationality and distortion have been given no independent definition or interpretation. BGS have simply stipulated that, in their model, a particular family of linear indifference curve is to be *called* ‘rational’.

BGS then specify ‘how salience distorts the valuation of a good’ (p. 810). The first step is to define a *salience function* which, for any choice set, for any good *k* in that set and for any attribute *j*, measures the degree to which the amount of attribute *j* ‘stands out’ (either as particularly high or particularly low – both are treated as sources of salience) relative to the average amount of that attribute in all goods in the choice set. The second step is to identify, for each good, which of the two attributes stands out more (as measured by the salience function). Thus, unless there is a tie, each good has a *salient attribute* – the attribute on which it stands out more.

For my purposes, it is sufficient to consider what BGS’s assumptions imply about salience when the choice set contains only two goods, with $p_1 > p_2$ and $q_1 > q_2$. These assumptions imply that if $q_1/q_2 > p_1/p_2$, quality is the salient attribute for both goods; if that inequality is reversed, price is the salient attribute for both goods. To get an intuitive feel for this property, think of the wine example. Good 1 is the French wine, with $p_1 = 20$ and $q_1 = 9$. Good 2 is the Austrian wine, with $p_1 = 10$ and $q_2 = 6$. Notice that $q_1/q_2 < p_1/p_2$. BGS’s assumptions imply that, in this case, the most salient feature of the French wine is its high price relative to the average price of the two wines; correspondingly, the most salient feature of the Austrian wine is its low price. But now suppose that the qualities of the wines are the same as before, but the prices are $p_1 = 50$ and $p_2 = 40$; now $q_1/q_2 > p_1/p_2$ (Suppose the choice is being made in a restaurant rather than a supermarket.) In this case, the most salient feature of the French wine is its high quality and the most salient feature of the Austrian wine is its low quality.

BGS's third step is to model the behaviour of a 'salient thinker' (i.e. a non-rational consumer) by 'distort[ing] the utility weights' that the consumer applies when evaluating goods. For the rational consumer, both attributes have a weight of 1 in the evaluation of every good. In contrast, when valuing any given good, the salient thinker uses a weight greater than 1 for its salient attribute and a weight less than 1 for its non-salient attribute (with the sum of the weights always equal to 2).

BGS apply this model to a wide range of consumer behaviour problems, using the general strategy of 'introducing salience-based valuation into a "rational" economic model' (p. 813). In these applications, they describe the effects of salience as 'distortions' of what would otherwise be 'rational' choices. All of this exemplifies the dualistic modelling strategy I described in Section 1. The behaviour of BGS's 'salient thinker' is determined by the interaction of two systems or processes – a set of context-independent latent preferences that are deemed to be rational, and a psychological mechanism which distorts these preferences. The choices of the salient thinker are determined by the distorted preferences, but the hypothetical choices of the rational consumer – that is, the consumer who acts on undistorted preferences – provide the normative benchmark. This is a model with an inner rational agent.

But what is the function of this benchmark in BGS's model? The essence of the model is that the relative weights of the two attributes differ according to which attribute is salient. But which attribute is salient for any given good in any given choice set depends only on the qualities and prices of the goods in that choice set, and these are defined independently of the consumer's latent preferences. Thus, any results that come about because of changes in relative attribute weights are independent of latent preferences. *The concept of latent preference serves no explanatory purpose.*

BGS's example of the two wines illustrates this point. In the story, the consumer chooses the lower-quality Austrian wine when the two wines are priced at $10 and $20, but the higher-quality French wine when the prices are $40 and $50. Leaving aside the possibility of perverse income effects, this pattern of choice is inconsistent with standard economic theory; but it has long been recognised as a common feature of human decision-making (see, for example, Savage, 1954, p. 103). It can be explained in various ways, for example by assuming diminishing sensitivity to changes in each attribute (Tversky and

Kahneman, 1991) or by assuming that expected prices act as reference points (Thaler, 1980); BGS provide a new explanation in terms of salience. In the example, adding $30 to the price of each wine switches the salient attribute from price to quality. (In general, adding any constant to the prices of each of two goods while keeping the qualities constant can cause a switch in the salient attribute; if there is a switch, it must be from price to quality. Thus any switch in choice must be from the lower-quality good to the higher-quality good.) But notice that all of this is true (in the model) irrespective of which good the consumer rationally prefers.

This does not mean that, in the world of the model, rational preferences are unobservable. Consider a decision problem in which there is only one good in the everyday sense of the word, but the consumer can choose whether or not to buy it. BGS represent this as a choice between $\langle p_1, q_1\rangle$ and $\langle p_2, q_2\rangle$, with $\langle p_2, q_2\rangle = \langle 0, 0\rangle$ representing 'not buying'. In this special case, BGS's preferred assumptions about the salience function imply that the two attributes are equally salient, and hence that the salient thinker's choices coincide with those of the rational consumer. Thus, rational preferences are revealed in the consumer's willingness to pay for individual goods in situations in which only one good is on offer. Remember, however, that up to this point, the concept of rationality has not been given any interpretation, except as the benchmark relative to which distortion is defined. So the only interpretation that can be given to the proposition that willingness to pay reveals rational preferences is that rational preferences *are defined by* willingness to pay.

To put this another way, the empirical content of the model is contained in the idea that choices between goods are influenced by the relative attention given to their attributes, and that more salient attributes are given more attention. A rational consumer is someone who always gives each attribute the right amount of attention. But what *is* the right amount of attention? In effect, BGS tell us that the right amount of attention to give each attribute is the attention that it is given in willingness-to-pay problems. But they do not explain what this statement means.

One possible reconstruction of the missing argument runs as follows. BGS are presupposing that the consumer has well-defined latent preferences between goods, defined as ⟨quality, price⟩ pairs, and that the latent utility of any good is independent of which other goods are in the choice set. This presupposition is essential for the rest of the reconstructed argument. For the cases that BGS's model is intended to represent, it is deemed an acceptable simplification to assume a weighted linear utility function, with the implication that the consumer has a context-independent utility weight α_Q / α_P for quality relative to price. Thus if (as in the wine example) his choices reveal implicit weights that are context-dependent, there must be some cases in which he chooses contrary to his latent preferences. If the qualitative pattern of context-dependence is consistent with a psychological theory of salience and attention, it is reasonable to infer that these are cases in which his rational judgement of the utility of the chosen good is distorted by salience effects deriving from comparisons between this good and other goods in the choice set. Thus, the best way to recover the consumer's latent preferences is to observe his choices in situations in which there is as little scope as possible for cross-good comparisons. Willingness-to-pay problems meet this requirement – at least in principle.

The 'in principle' qualification is needed because BGS extend their basic model to allow the salience of an attribute to depend not only on the content of the choice set, but also on 'alternatives that the decision maker expects to find in the current choice setting', and hence on the consumer's expectations about prices (p. 820). Thus, this method of eliciting latent preferences requires a setting in which 'a good is evaluated in isolation and without price expectations'. BGS suggest that such settings can be created in 'lab experiments' (p. 828). In the light of decades of attempts to elicit willingness-to-pay valuations in experiments and surveys, this suggestion seems extraordinarily optimistic. Responses to willingness-to-pay and willingness-to-accept questions are known to be influenced by many kinds of irrelevant cues which draw attention to particular answers (Parducci, 1965; Slovic and Lichtenstein, 1968; Johnson and Schkade, 1989; Ariely et al. 2003). For example, if the elicitation exercise begins with a question of the form 'Would you be willing to pay $*x*?', final responses are pulled towards $*x*; if respondents are asked to pick a point on a scale of possible values, responses are pulled towards the middle of the scale. These 'anchoring' and 'range/frequency' effects are particularly strong when (as in stated preference studies which try to elicit valuations for non-marketed goods, such as changes in environmental quality) there is no customary price that the respondent can use as a benchmark. A natural interpretation of this evidence is that people find it very difficult to give a monetary valuation of any good *in isolation* and that, when required to do so, they unconsciously search for comparators and reference points.

So it is far from self-evident that individuals have well-defined context-independent latent preferences, ready to be elicited by economists. Since latent preferences play no role in BGS's explanation of actual choices, we have been given no reason to think that the mental processes that lie behind these choices make use of any such construct. But it is only by assuming the existence of latent preferences that the concepts of ‘rationality’ and ‘distortion’ can be given any independent meaning.

4. A psychological model of attention

My second example is a seminal contribution to the psychology of decision-making under uncertainty – *decision field theory,* as proposed by Busemeyer and Townsend (hereafter ‘BT’; 1993). BT’s aim is ‘to understand the motivational and cognitive mechanisms that guide the deliberation process involved in decisions under uncertainty’. They are particularly concerned with explaining two ‘unavoidable facts about human decision making’ – that the preferences of a given individual over given pairs of alternatives are subject to stochastic variation, and that the amount of time spent making a decision influences the final choice (pp. 432–435). Thus, they need a model in which deliberation about what to choose is a process that occurs over time and includes some random element.

BT’s basic model is of an individual who has to choose between two *actions* in a situation of uncertainty. Uncertainty is represented by a set of alternative *events,* one and only one of which will occur. An action is defined by the *payoff* that will occur in each event if that action is chosen. Payoffs are implicitly assumed to be measured on a ratio scale and can be positive, zero or negative. BT assume the existence of a utility function which assigns a real value $u(x)$ to every payoff x. However, the individual is not assumed to attach objective probabilities to events. BT say that they are dealing with decisions under *uncertainty* (as opposed to risk), defined as problems in which ‘the decision maker must learn and infer the event probabilities from past experience’ (p. 436).

Notice the formal similarities between this problem and the one studied by BGS. BT’s ‘actions’ and ‘events’ are respectively analogous with BGS’s ‘goods’ and ‘attributes’. The ‘payoffs’ of actions in events are analogous with the ‘amounts’ of attributes that goods possess. BGS’s model does not have an explicit analogue of BT’s utility function for payoffs, but that is only because BGS assume that utility is linear in amounts of attributes.[5] In BT’s

[5] One disanalogy between the problems should be pointed out. BT’s utility function is defined on payoffs, independently of the events in which they occur, and so *event-independent* utility measures are treated as inputs to the deliberation process. On my reading, BGS implicitly assume *attribute-specific* utilities as the analogous inputs.

model the individual's problem is to make trade-offs between payoffs that occur in different events; in BGS's model, it is to make trade-offs between amounts of different attributes.

It may help to keep in mind a concrete example of a decision problem to which BT's model might be applied. Consider Jane, who will be working in some city for a fixed period and has to decide whether to buy a house or to rent one. She knows the current purchase and rental prices of property but is uncertain about how these prices will change over the period. If property prices rise, she will gain by buying rather than renting; if they fall, she will lose. She cannot assign objective probabilities to these events. (In fact, no one can: if she consults supposed experts, she will find that their judgements differ.)   In this problem, the actions are 'buy' and 'rent' and the events are alternative rates of change in property prices.

BT present decision field theory as a succession of amendments to *deterministic subjective expected utility theory*, interpreted as a decision rule which assigns a weight to every event (normalised so that the weights sum to 1) and chooses whichever action has the higher weighted average utility. Expected utility theorists normally interpret each of these weights as a subjective probability, but BT offer a different interpretation, saying: 'From a cognitive view, this weight reflects the *amount of attention* given to [the relevant event] on each presentation of the choice problem' (p. 436; italics in original). Thus, BT's model, like BGS's, is one in which decisions depend on the distribution of the decision-maker's attention (between events or between attributes).

In decision field theory, deliberation is a process that occurs over time. At any given moment during this process, there is a *preference state* measured on a real-valued scale; positive values represent strength of preference in favour of one of the actions, negative values represent strength of preference in favour of the other. Deliberation begins with an initial preference state. In 'neutral' versions of the theory, the initial state is zero, but BT allow the possibility that the initial state is 'biased by past experience' in the direction of the individual's decisions in similar previous problems (p. 441). This mechanism has the effect of reducing decision times and increasing the stability of choice in familiar problems.

Deliberation is represented as *sequential sampling* of events. Each time an event is sampled, the utility difference between the two actions in that event is registered, and the preference state is updated in the direction of the action with the higher utility. Deliberation ends when the preference state crosses a pre-determined upper or lower *threshold*; the action that is preferred in this state is then chosen. Sampling an event is interpreted as *attending to* its payoffs: 'The basic idea is that attention may switch from one event to another within a single choice trial' (p. 438).

Clearly, the probability that a given action is chosen depends on (among other things) the probabilities with which the different events are sampled. Under certain neutral assumptions (including that the sampling probability for each event remains constant during the process, that the initial preference state is zero, and that the upper and lower thresholds have the same absolute value), the action that is more likely to be chosen is the one with the higher weighted average utility when each event is weighted by the probability that it is sampled at each stage of the process. In other words, under these assumptions it is *as if* the individual has a subjective probability for each event and is more likely to choose the action with the higher subjective expected utility; but the as-if probability of any event is actually a measure of the individual's propensity to attend to it in the deliberation process. On a strict reading of BT, whether this as-if probability can be interpreted as the individual's subjective judgement of the likelihood of the event itself is left open. One might say that, in leaving this question open, BT are working in the spirit of Savage's (1954) subjectivist interpretation of probability as a property of an individual's preferences over actions, as revealed in her decisions.

In its most general form, decision field theory does not impose these neutral assumptions, and so does not necessarily generate decisions that can be rationalised by a stochastic form of subjective expected utility theory. But, given the utility function, the initial preference state, the decision thresholds and a full specification of the mechanism which determines the distribution of the individual's attention, BT's model generates stochastic decisions and associated decision times. With one exception, it does so without using any concept of 'correctness' in decisions.

The exception appears in BT's discussion of the implications of alternative values of the threshold, on the simplifying assumption that the upper and lower thresholds have the same absolute value θ. In this discussion, BT define the *correct* action as the action that produces the higher subjective expected utility. They then say:

> … the threshold criterion θ controls speed–accuracy or cost–benefit trade-offs in decision making. One the one hand, if the cost of prolonging the decision is low or the cost of making an incorrect decision is high, then a high threshold is selected. On the other hand, if the cost of prolonging the decision is high or the cost of making an incorrect decision is low, then a low threshold is selected. (p. 440)

In a footnote to this passage, BT discuss a possible amendment to their model, according to which the value of θ decreases over the deliberation period. Since the fact that the threshold has not been crossed after a long time is evidence that the difference in attention-weighted utility between the actions is relatively low, this amendment would implement what might be called a speed–accuracy trade-off by means of a simple and well-defined psychological mechanism. If this is all that BT have in mind in the quoted passage, nothing much hangs on their definition of 'correctness'. Nevertheless, that definition is question-begging. For BT, subjective expected utility is merely a construct which, under certain assumptions, can be read off from the decisions produced by the sequential sampling process; the as-if probabilities used in the definition of this construct are determined by the distribution of attention. It is not clear why the individual's propensities to attend to the different events should determine which action is deemed to be the correct choice.

Think about Jane choosing between buying a house and renting one. If she deliberates in the way described by BT's model, her attention will switch in a random fashion between thinking about a rising property market (and about the corresponding benefits of buying) and thinking about a falling property market (and about the corresponding benefits of renting). Suppose that, if she deliberates for a long time and with many switches of attention, she can be expected to spend 60 per cent of the deliberation period thinking about a rising market and 40 per cent of the period thinking about a falling market. How can that fact make the correct choice for Jane be the one that has the higher expected utility when the probabilities of the two events are set at 0.6 and 0.4?

One possible answer is that BT's concept of 'correct' choice is not intended to be normative, in the sense of saying what the individual *ought* to choose; rather, it is an empirical concept, referring to the long-run tendency of deliberation. (It would not sound so odd to say that Jane has a *latent preference* for the action that she would be more likely to choose after long deliberation.)

BT may also be thinking of possible extensions of their theory which could close the gap between the normative and empirical concepts of correctness. Recall that in their definition of 'uncertainty', they refer to *the* event probabilities that the decision-maker has to learn from experience. It would not be inconsistent with this definition to assume the *existence* of event probabilities, perhaps defined as relative frequencies in a (possibly hypothetical) series of exactly repeated trials. Perhaps BT are entertaining the hypothesis that, if an individual faces exactly the same decision problem many times, the distribution of her attention between events converges to the corresponding distribution of event probabilities, irrespective of contextual factors. If this *attention hypothesis* were true (and given other neutral assumptions), the long-run tendency of repeated decision-making would be towards a state in which the action that was more likely to be chosen was the one with the higher weighted average utility when each event is weighted by its 'objective' probability. One might call this action 'latently preferred' (or the 'correct' choice) in an empirical sense. If one believed that expected utility theory was grounded on compelling principles of rationality, one might also call that action 'correct' in a normative sense. And so, if the attention hypothesis were confirmed, one might claim that decision field theory isolates the psychological substrate of context-independent latent preferences of just the kind that behavioural welfare economics needs.

But there are some very big 'if's here. Notice in particular that the argument sketched in the previous paragraph depends on assumptions that imply that, if the same decision problem is repeated many times, any systematic context-dependence effects gradually disappear. If it really were the case that the choices of experienced decision-makers reliably revealed context-independent preferences, most behavioural economists would probably agree that the preferences to be used in normative analysis should be those that individuals reveal after having sufficient experience of relevant choice problems. But the truth is that, after a quarter of a century of experimental investigation of the influence of experience on decision 'anomalies', the only general conclusion that can be drawn is that some but not all anomalies seem to decay with some but not all kinds of experience.[6] I think we have to accept that context-dependent choice is not just a symptom of inexperience.

Indeed, one might think that the fundamental principles of decision field theory provide reasons for expecting context-dependence to be a pervasive and persistent feature of human decision-making. If the distribution of an individual's attention between alternative events or different attributes is a crucial determinant of her decisions, any 'irrelevant' factor which influences the distribution of attention will be capable of inducing context-dependent choices. It does not seem at all self-evident that these influences will become less powerful or less effective as a decision-maker gains experience.

[6] This literature is too large and diverse to be usefully reviewed in a philosophically-oriented paper. Loomes, Starmer and Sugden (2010) report one experiment which found mixed results, and refer to other relevant papers.

If context-dependence is a systematic and persistent consequence of psychological mechanisms that control the distribution of attention, behavioural welfare economics has to face the question to which BGS's model provided no satisfactory answer: How can we identify context-independent latent preferences? As an explanation of how attention influences choice, decision field theory is much deeper and more convincing than BGS's economic model; but it does not answer that question. What it does do is to help us understand why the presupposition of the question is mistaken. If context-independent latent preferences play no role in psychological explanations of actual deliberation or actual choice, we should not expect psychology to tell us how to identify them.

5. Discussion

The aim of behavioural welfare economics, as I understand it, is to show how welfare judgements or public policy decisions can respect each individual's own subjective preferences, as revealed in her choices after the effects of psychologically-induced errors have been controlled for. My purpose in discussing these two models of attention was to explore whether a psychological analysis of decision-making as mental processing might allow an empirical distinction to be made between latent preferences and error. These models are interesting because they represent attention-based mental processes that can induce context-dependent choices, and because their authors – particularly the authors of the model that belongs to behavioural economics – make use of concepts of 'rational' or 'correct' latent preference.

I have argued that these concepts serve no explanatory purpose. In these models, individuals' decisions depend on the relative attention given to different attributes or events, allowing context-dependent choices to be explained by causal factors that impact on the mental processes that control the distribution of attention. Of course, if one chooses to define any particular preference as 'correct', there is a correspondingly 'correct' distribution of attention. And given any such definition of correctness, there is a corresponding definition of 'error', namely that an error occurs when an incorrect choice is made; the cause of the error is an incorrect distribution of attention. One might choose to call this causal mechanism a 'bias' or 'distortion' of correct reasoning. But none of this is any help in determining *which* preferences are latent in the individual and which are not.

If behavioural welfare economics is to succeed in its aim, it has to be able to identify some mode of reasoning or mental processing which, in some well-defined hypothetical situation, would lead the individual to reveal context-independent latent preferences; it must have some defensible criterion for defining errors in reasoning; and it must provide good reasons for thinking that this situation is one in which such errors are particularly unlikely to occur. I submit that it has not found any way of doing this, and that the prospects of success are poor. The root of the problem, I believe, is that when economists (and indeed many philosophers, and perhaps even some psychologists) think about human agency, they find it hard to avoid using a mental model in which humans are ultimately rational beings. This model may recognise that humans can hold irrational beliefs and make irrational decisions, but at some deep level, irrationality is understood as the product of mistakes. These mistakes must be defined relative to some 'true' preferences – the preferences of the human individual's 'true self'. This is the model of the inner rational agent. We need to recognise that this model is pre-scientific. To questions about the role of latent preferences, the best answer is the one that, in another context, Laplace gave to Napoleon: I had no need of that hypothesis.

References

Ariely, Dan, George Loewenstein and Drazen Prelec (2003). Coherent arbitrariness: stable demand curves without stable preferences. *Quarterly Journal of Economics* 118: 73–105.

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.

Broome, John (2013). *Rationality Through Reasoning*. Oxford: Wiley Blackwell.

Busemeyer, Jerome and James Townsend (1993). Decision field theory: a dynamic–cognitive approach to decision making in an uncertain environment. *Psychological Review* 100: 432–459.

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 (2014). Common reasoning in games: a Lewisian analysis of common knowledge of rationality. *Economics and Philosophy* 30 (2014): 285–329.

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

Infante, Gerardo, Guilhem Lecouteux and Robert Sugden (2015a). Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. Forthcoming in *Journal of Economic Methodology*.

Infante, Gerardo, Guilhem Lecouteux and Robert Sugden (2015b). 'On the Econ within': a reply to Daniel Hausman. Forthcoming in *Journal of Economic Methodology*.

Johnson, Eric and David Schkade (1989). Bias in utility assessments: further evidence and explanations. *Management Science* 35: 406–424.

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.

Loomes, Graham, Chris Starmer and Robert Sugden (2010). Preference reversals and disparities between willingness to pay and willingness to accept in repeated markets. *Journal of Economic Psychology* 31: 374–387.

McQuillin, Ben and Robert Sugden (2012). How the market responds to dynamically inconsistent preferences. *Social Choice and Welfare* 38: 617–634.

Parducci, Allen (1965). Category judgment: a range-frequency model. *Psychological Review* 72: 407–418.

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.

Slovic, Paul and Sarah Lichtenstein (1968). Relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology* 78: 1–18.

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 (2007). The value of opportunities over time when preferences are unstable. *Social Choice and Welfare* 29: 665–682.

Taylor, Shelley and Suzanne Thomson (1982). Stalking the elusive vividness effect. *Psychological Review* 89: 155–181.

Thaler, Richard (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization* 1: 39–60.

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