International Review of Economics (2021) 68:5–19
https://doi.org/10.1007/s12232-020-00356-8

RESEARCH ARTICLE

Normative economics without preferences

Robert Sugden[1]

Received: 25 May 2020 / Accepted: 27 May 2020 / Published online: 23 July 2020

Keywords Community of advantage · Behavioural welfare economics · Reconciliation problem · Paretian liberal

JEL Classification D60 · D91

I am pleased and honoured that *The Community of Advantage* is the subject of this collection of insightful papers (Sugden 2018). It is particularly appropriate that the symposium is published in what is coming to be known as the *Journal of Civil Economy*. The understanding of economic life as cooperation for mutual benefit that I am trying to express through the phrase 'community of advantage' is in the same spirit as 'civil economy', the name that Antonio Genovesi—the Neapolitan who was the world's first university professor of economics—wanted to give our discipline.

In this paper, I explain how the ideas in my book fit into the broad landscapes of normative and behavioural economics. I have the sense that some of my fellow behavioural economists see my approach to normative issues as not so much contractarian (which it professes to be) as contrarian. Undoubtedly, I deviate from what has become the mainstream position of behavioural economics. But that may reflect the fact that, although I have been a practitioner of behavioural economics since the pioneer era of the 1980s, I have worked at the interface of economics and philosophy for even longer. Much of my work as a philosophical economist has been concerned with concepts of liberty and opportunity. I will try to explain how these investigations have influenced my understanding of the normative implications of the findings of behavioural economics.

✉ Robert Sugden
r.sugden@uea.ac.uk

[1] School of Economics and Centre for Behavioural and Experimental Social Science, University of East Anglia, University Plain, Norwich NR4 7TJ, UK

1 What is behavioural economics?

The title of my book is taken from a passage in John Stuart Mill's *Principles of Political Economy*, in which Mill describes international trade—and, by implication, the market system—as a 'community of advantage' (1871/1909, Book 2, Chapter 17, Sect. 5). My subtitle, *A Behavioural Economist's Defence of the Market*, was chosen to suggest that the ideas in the book would be controversial. I considered an alternative, 'A behavioural defence of the liberal tradition of economics', which would have been intellectually more precise but less arresting. What I wanted to convey was my opposition to the fashionable view that the findings of behavioural economics undermine a tradition of economics that goes back to Mill, Adam Smith and Genovesi, in which the market system is viewed favourably (which is not to say uncritically). I wanted to say that it is possible to be both a behavioural economist and a liberal economist.

What is a behavioural economist? The term 'behavioural economics' seems to have been introduced in the 1990s in a re-branding exercise. The aim was to separate two programmes of experimental research which, up to then, had shared the name 'experimental economics'. One programme, particularly associated with Vernon Smith, was to keep the name 'experimental economics'. The other, particularly associated with Daniel Kahneman and Amos Tversky, was to become 'behavioural economics'. In a paper explaining this separation, George Loewenstein (1999: F25) defines a behavioural economist as 'an economist who brings psychological insights to bear on economic phenomena'. I think that is an apt definition. It is in that sense that I can claim to be a behavioural economist.

However, I recall a conversation with Kahneman in the early 2000s in which I asked him how he would define behavioural economics. He wanted to exclude Maurice Allais's (1953) discovery of the common consequence and common ratio effects from the canon of behavioural economics, even though his and Tversky's experimental replications of those effects were two of the central exhibits of their 'prospect theory' paper—the paper that most behavioural economists regard as a founding text (Kahneman and Tversky 1979). Personally, I have to confess, I was more concerned that Kahneman classified Graham Loomes's and my regret theory as non-behavioural (Loomes and Sugden 1982). His explanation for these exclusions was that theories *of rational choice* were not behavioural. Allais had used the common consequence and common ratio effects to justify a theory of rational choice that did not include the independence axiom. Loomes and I had explicitly presented regret theory as 'an alternative theory of rational choice under uncertainty'. My response (which did not persuade Kahneman) was that prospect theory and regret theory had both been constructed by using expected utility theory—the standard theory of rational decision-making—as a template and then adding particular psychological mechanisms (loss aversion and probability weighting in one case, regret in the other).

I now see that there *is* a significant difference between the way that rationality is treated in those two papers. After describing the differences between prospect theory and expected utility theory, Kahneman and Tversky (1979, p. 277) conclude:

> These departures from expected utility theory must lead to normatively unacceptable consequences, such as inconsistencies, intransitivities, and violations

> of dominance. Such anomalies of preference are normally corrected by the decision maker when he realizes that his preferences are inconsistent, intransitive, or inadmissible. In many situations, however, the decision maker does not have the opportunity to discover that his preferences could violate decision rules that he wishes to obey. In these circumstances the anomalies implied by prospect theory are expected to occur.

For many years, I read this passage merely as a tactical concession to rational choice theorists, suspecting that it might have been written to accommodate a difficult referee. Taken at face value, however, it says that prospect theory is *a theory of error*, and that error can be defined relative to the *correct* theory of rational choice, namely expected utility theory.

In contrast, Loomes and I claimed that regret theory was *a* (not *the*) theory of rational choice: it was a model of a psychologically explicable form of behaviour that could not reasonably be deemed *ir*rational. We presented this model as a counter-example to the claim that transitivity, independence and respect for first-order stochastic dominance were necessary properties of rational preferences. Around the same time, we developed a parallel theory of 'disappointment', in which preferences that we claimed were rational contravened the sure-thing principle—just about the only major rational-choice principle that regret theory did *not* violate (Loomes and Sugden 1986). We thought that, by bringing psychological insights about regret and disappointment to bear on economic phenomena, we had shown the implausibility of conventional claims about the normative status of expected utility theory—including those made by Kahneman and Tversky in the passage I have quoted. This is certainly an enterprise of behavioural economics according to Loewenstein's definition. In any case, it would be odd to treat acceptance of the normative status of a non-psychological theory as a defining characteristic of the branch of economics that makes use of psychology.

For the first two decades after the publication of Kahneman and Tversky's 'prospect theory' paper, behavioural economics was almost entirely concerned with investigating and explaining facts of human behaviour; it did not engage with normative issues. Whether the regularities of behaviour that were being observed were errors relative to some true concept of rationality was not a major issue. But this changed in the early 2000s when attention turned to what I have called the *reconciliation problem*—the problem of reconciling normative economics with the empirical findings of behavioural economics.

2 Neoclassical welfare economics and social choice theory

From the 1930s to at least the 1980s, there was a clear consensus among economists about the theoretical framework within which normative analysis was conducted. This framework rested on a very parsimonious conceptual scheme, essentially as follows.

A *society* consists of a set of *n individuals*. A *social state* is a possible description of society. This may be a single description (for example, 'the socialist party

forms the government') or an *n*-tuple of individual-specific descriptions (for example, specifying each individual's bundle of consumption goods). For each individual, there is a binary relation of *preference* over the set of all relevant social states. Preferences are complete (that is, the individual can rank each pair of social states in terms of preference or indifference) and *integrated* (that is, non-stochastic, context-dependent and internally consistent in the sense defined by axioms such as transitivity and the sure-thing principle). An individual's preferences are revealed in the *choices* she makes from *opportunity sets* of feasible objects. Normative economics is concerned with the *social ranking* of social states. The *Pareto principle* prescribes that if all individuals weakly prefer some social state x to another social state y, then x is socially ranked at least as highly as y; if in addition at least one individual strictly prefers x to y, then x is ranked strictly above y. Most applications of normative economics accepted this principle.

However, this conceptual scheme was open to a range of different interpretations, each of which was compatible with the standard neoclassical theory of individual behaviour. It was possible, and indeed quite common, for economists to use the scheme without saying how they interpreted it.

Crucially, the concept of preference could be interpreted in different ways. On some readings of revealed preference theory, preference *means* choice (subject to some qualifications about indifference): to say that a person prefers x to y is to say that if she faces the opportunity set $\{x, y\}$, she *in fact* chooses x. On another interpretation, a preference is a mental state that *tends to cause* choice: to say that a person prefers x to y is to say that she is psychologically disposed to choose x from $\{x, y\}$. On a third interpretation, a preference is an individual's subjective judgement about her welfare: to say that a person prefers x to y is to say that, in her judgement, x would be better for her than y. If individuals are self-interested and rational in the sense of neoclassical theory, these interpretations are observationally equivalent. The person's judgement that x is better for her than y provides her with good reason both to be disposed to choose x rather than y and to act on that disposition.

The concept of a social ranking was equally open to different interpretations. On one interpretation, it is a normative judgement made from a neutral viewpoint by a 'social planner' or 'ethical observer': to say that x has a higher social ranking than y is to say that, as judged by a neutral observer, x is better for society than y. On another interpretation, it is an aggregation of individuals' preferences: to say that x is ranked above y is to say, roughly speaking, that there is a greater weight of preference on the x side than on the y side. (Needless to say, the meaning of such an aggregation depends on how the concept of preference is interpreted.) On a third interpretation, a social ranking is the actual result of some (or perhaps of a recommended) collective choice mechanism: to say that x is ranked above y by some specific mechanism is to say that that mechanism would choose x from $\{x, y\}$.

The open-endedness of the concepts of preference and social ranking left room for many different interpretations of the Pareto principle. If a social ranking is a 'better for society' judgement made by a neutral observer, the Pareto principle might be a normative principle prescribing respect for individuals' choices, dispositions or subjective judgements. Or it might be a normative principle requiring the observer to take account of each individual's welfare, defined in some objective

sense, combined with the assumption that individuals in fact tend to prefer what is objectively good for them. If a social ranking is an aggregation of preference, the Pareto principle might be a minimal, non-normative aggregation principle. (If some component of an aggregate increases in quantity and if no component decreases, the aggregate increases.) If a social ranking is the result of a recommended collective choice mechanism, the Pareto principle might be a minimal normative condition of democracy. (Voting rules should be such that, if a social decision has to be made between x and y, and if no one votes for y and at least one person votes for x, then x should be declared the winner.)

The variety of interpretations of the Pareto principle allowed a corresponding variety of interpretations of the First Fundamental Theorem of welfare economics—the theorem that every competitive equilibrium is Pareto-efficient. In particular, it could be interpreted as a conclusion about how competitive markets contribute to social welfare, or as a conclusion about the tendency of such markets to respect consumer sovereignty by responding to individuals' actual demands. Thus, economists could agree that the First Fundamental Theorem picked out a desirable feature of markets without actually agreeing about what that feature was.

Some of the tensions in economists' apparent consensus about normative analysis began to emerge in the 1970s and early 1980s, when social choice theory was one of the most fashionable areas of economic theory and a topic of interest to many philosophers. But developments in behavioural economics revealed more fundamental problems. The long-standing consensus was based on a shared sense that choice, preference and welfare were closely related. But a recurring finding of behavioural economics is that individuals' choices are sensitive to features of the 'context' or 'framing' of decision problems that seem to have no relevance to individuals' interests or well-being. The implication is that individuals often do not have the kind of preferences that economists have traditionally assumed when treating preference as the central concept in their normative analyses. Something in the consensus position has to be given up. Choosing what that should be, and so finding a way forward for normative economics is the reconciliation problem.

3 Non-meddlesome preferences and the problem of the Paretian liberal

My approach to this problem builds on what I learned through my involvement in debates about the interpretation of the concepts of 'individual preference' and 'social ranking' that took place among social choice theorists in the 1970s and 1980s. A theorem due to Sen (1970), 'The impossibility of a Paretian liberal', crystallised some of the central issues in these debates and was the subject of my first work in the philosophy of normative economics. Let me explain.[1]

[1] This section is based on Sugden (1985), which provides more detail about these debates. I have re-used some text from that paper.

For Sen, this theorem was part of a research programme grounded on the belief that normative economics, as practised at the time, had too narrow an 'informational base'. Sen shared the consensus view that normative economics was about social rankings of social states, but challenged the 'welfarist' assumption that, in arriving at judgements about social rankings, the only relevant information was information about individuals' preferences. His theorem is intended to show that the welfarist approach cannot take account of widely shared moral intuitions about the value of individual liberty.

Since Sen is presenting an impossibility result, he defines what he sees as a minimal condition of social respect for liberty. His condition of *minimal liberty* is that there should be some protected personal sphere for each of at least two individuals in society. Which individuals should have this privilege, and what should belong to their personal spheres, is left open, so on the face of it this is an extremely weak requirement. A personal sphere is to be understood as a nonempty set of pairs of social states. If some pair of social states $\{x, y\}$ is deemed to belong to person i's sphere and if i prefers x to y, then x must be socially ranked above y. Sen (1982, p. 286) says that the minimal liberty condition is intended 'to permit each individual the freedom to determine at least one social choice, for example, having his own walls pink rather than white, other things remaining the same for him and the rest of the society'. The theorem shows that, for any society and for any specification of personal spheres that satisfies the minimal liberty condition, there is some profile of individual preferences such that *either* the Pareto principle is violated *or* the social ranking contains a cycle.

In my contributions to the debate about the theorem, I gave prominence to the following example, based on one originally presented by Gibbard (1974). In a strict sense, it is not a proof of Sen's theorem, but it illustrates the theorem's underlying normative logic. The example should be read in the context of the marriage norms of the 1970s. There are three unmarried individuals, Annie, Bill and Charlie. Three alternative social states are feasible: that no one marries anyone, x; that Annie marries Bill, Charlie remaining single, y; and that Annie marries Charlie, Bill remaining single, z. On any normal conception of liberty with respect to marriage, every individual has the right to stay single if he or she so chooses, and any adult man and adult woman (if not close relatives and not already married to anyone else) have the right to marry if they both choose to do so. The first of these propositions fits naturally into Sen's conception of the personal sphere. Consider any two social states that differ only in respect of whether a particular man is married to a particular woman—for example, x and y, which differ only in respect to the marriage or non-marriage of Annie and Bill. It would be in the spirit of Sen's characterisation of liberty to say that either Annie's preferring x to y or Bill's preferring x to y is a sufficient condition for x to be socially ranked above y. Indeed, Sen (1982, pp. 299–302) seems to endorse this formulation in a discussion of liberty in relation to marriage. An analogous formulation of the second proposition would be to say that if the only difference between two states (such as x and y in the example) is whether a particular man and woman are married to one another, and if both of them prefer being married to one another to remaining single, then this is also the social ranking.

As I like to tell the story, Annie and Bill are long-standing friends. Annie would like the two of them to marry. Bill would prefer the relationship to continue without any commitment, but would rather marry Annie than lose her altogether. Charlie admires Annie and would like to marry her. Annie is fond of Charlie and would settle for him if it was clear that Bill would not marry her. Representing this familiar triangle in terms of transitive individual preferences: Annie prefers y to z and z to x; Bill prefers x to y and y to z; and Charlie prefers z to x. In one possible version of the story, Charlie (perhaps modelling himself on the hero of Dickens's *Tale of Two Cities*) self-sacrificingly prefers y to z. This leads to a variant of Sen's impossibility result. Socially, z is ranked above x (because this is the shared preference of Annie and Charlie) and x is ranked above y (because this is Bill's preference). But if the social ranking is also required to satisfy the Pareto principle, y is ranked above z, producing a cycle.

However, I would rather have Charlie believing that he is the right man for Annie, and so I tell the story with Charlie preferring z to x and being indifferent between x and y. Forget about social rankings for the moment and ask what would happen if these preferences were common knowledge and if Bill and Charlie were both free to make proposals of marriage to Annie. Charlie has nothing to lose by proposing. If Bill expects Charlie to propose, the best he can do it to propose too. Having received both proposals, Annie will accept Bill. (If the story were set on 29 February, when by tradition women propose to men, Annie's best strategy would be to propose to Bill, making it clear that if he rejected the proposal she would propose to Charlie. The result would be the same.) So, by permitting marriages to be arranged by a propose-and-accept rule, society allows y (Annie marries Bill) to come about when x (no one marries anyone) is feasible. But Bill prefers x to y. If we accept Sen's characterisation of liberty, respect for Bill's liberty requires that x is socially ranked above y: we must conclude that the workings of the propose-and-accept rule have violated Bill's protected personal sphere. Now the paradox is not an inconsistency between liberty and the Pareto principle; it is that Sen's characterisation of liberty is telling us that, if society is to protect individual liberty, it cannot allow marriages to be formed by free choice and mutual consent. Another way of expressing the sense of paradox is to say that the liberty that (supposedly) has not been respected is Bill's freedom not to marry Annie, but Bill has in fact chosen to marry her. Has he invaded his own personal sphere?

Sen's impossibility result initiated a huge literature in which social choice theorists tried to find coherent ways of representing individual liberty. Most contributions to this literature framed this problem in the same way that Sen had done—as a problem of arriving at a social ranking of social states, using information about individuals' preferences over those states. For many writers, the starting point was the recognition that that Sen's result (along with a whole range of related puzzles) is possible only if some individuals have strict preferences over pairs of states that belong to other individuals' personal spheres. (In the first version of my example, $\{x, z\}$ belongs to Annie's and Charlie's spheres, but Bill prefers x to z; $\{x, y\}$ belongs to Annie's and Bill's spheres, but Charlie prefers y to x.) Such preferences were often characterised as 'meddlesome'. One response was to look for ways of editing (or 'laundering') individuals' preferences so as to remove elements of meddlesomeness

before using them as inputs to the formation of social rankings. Sen's liberty principle might then be retained by applying the Pareto principle to laundered rather than actual preferences. Another response was to argue that if, in some domain of life, an individual is sufficiently meddlesome, his or her right to a protected personal sphere should be deemed to have lapsed. The Pareto principle might then be retained by restricting the scope of the liberty principle. (Compare the popular view that the rights of free speech and assembly should not be given to anti-democratic organisations.) A third response, in the spirit of Sen's critique of the overly narrow informational base of welfare economics, was to argue that the social ranking should take account of the intentions and motivations that lie behind meddlesome preferences. (Does Bill's preference for x over z reflect spite, jealousy, or just a desire to maintain a long-standing friendship?) Increasingly complex—not to say baroque—formulations of liberty were proposed, discussed, and found to generate yet more paradoxes.

An initially small minority of theorists, of whom I was one, argued that the source of the problem was the presupposition that respect for liberty should be represented as a relationship between individual preferences and social rankings (Nozick 1974, pp. 164–166; Sugden 1978, 1985; Gärdenfors 1981). Although we never managed to persuade Sen, we did convince three of the leading social choice theorists of the time (Gaertner et al. 1992).

Our argument was that liberty is a property of the procedure that determines which social states come about, and so is a relationship between individuals' *choices* (not their preferences) and social *outcomes* (not rankings). As Peter Gärdenfors pointed out, procedural properties of social choice can be represented using the theoretical framework of *game forms*—that is, games with the same properties as those of classical game theory, except that outcomes are described in physical terms, rather than as profiles of players' utilities. In this conceptual scheme—sometimes called the *game form approach*—individuals' preferences play no role; 'social' judgements are not about the outcomes of a society's decision-making procedures, but about the procedures themselves. In the case of marriage, respect for individual liberty makes certain requirements of *the procedure by which marriages come about*. Specifically, the marriage of any single and adult man and woman should occur if and only if they both choose to marry. Either version of the propose-and-accept rule satisfies this requirement. In the example, the social state that comes about—Annie marries Bill, Charlie remaining single—does so through the workings of a procedure that respects individual liberty. That is all that needs to be said. Viewed in the perspective of liberty, Annie, Bill and Charlie's preferences, intentions and motivations have no bearing on the case.

In some respects, the game form approach is aligned with a way of thinking that became important in moral and political philosophy around the same time as the debates I have described. This was the idea that each individual's well-being or 'advantage' should be assessed in terms of the opportunities from which she can choose, rather than in terms of the outcomes she experiences. This idea is central to John Rawls's theory of justice, in which holdings of 'primary goods'—things that 'normally have a use whatever a person's rational plan of life'—are the metric of advantage (Rawls 1971, p. 62). It is also central to Sen's (1985) theory of *capability*, in which a person's advantage is assessed in terms of the set of combinations of

'beings and doings' from which she can choose. This new interest in opportunity prompted a line of research whose aim was to find ways of measuring the extent or value of opportunity that is provided by a person's opportunity set. As a contributor to this research programme, I argued that the extent of an individual's opportunity should be assessed without reference to her actual preferences, but only in terms of what is 'normally' or 'reasonably' preferred by people in general (Jones and Sugden 1982; Sugden 1998).

However, the example of Annie, Bill and Charlie illustrates the limitations of an approach to normative analysis that assesses each individual's opportunity set independently of those of other people and then looks for a fair or equal distribution of opportunity. The problem is that the contents of each individual's opportunity set depend on the choices that other individuals make from theirs. Given the propose-and-accept rules for forming marriages, an individual's opportunities to marry are not simple properties of his or her opportunity set. For example, Charlie's opportunity to marry Annie is an opportunity for him to marry her *if she also chooses to marry him*. Given Annie's preference for Bill, the opportunity set that Charlie faces *as an individual* is less rich than Bill's. Nevertheless (and contrary to the position that seems to be taken by adherents of the 'involuntary celibacy' movement), I think we should say that the propose-and-accept rules treat Bill and Charlie equally. The interrelatedness of opportunity sets is crucial for the analysis of market opportunities in *The Community of Advantage*.

But let me take a step back and consider some of the general methodological features of the controversy about how to formulate principles of individual liberty. The original problem was to give a formal representation of an intuitive normative principle—the principle that society ought to respect individuals' personal spheres. Social choice theorists tackled this problem within a pre-existing conceptual framework in which normative judgements are expressed in social rankings and social rankings depend on individuals' preferences. Accordingly, they tried to represent the intuitive principle as a particular kind of relationship between preferences and social rankings. It turned out that *this representation* captured the intuition only if individuals' preferences had certain properties of non-meddlesomeness. Many theorists seem to have concluded from this that *the principle itself* was applicable only if preferences had those properties. They responded by proposing analyses which *either* 'corrected' individuals' preferences to fit the representation, *or* disqualified non-fitting preferences as normatively unacceptable, *or* required investigations into individuals' intentions or motivations as a means of deciding whether particular preferences should be respected. Intuitively, none of these proposals seems compatible with the simplicity of the original principle—the principle that in certain personal matters, individuals should be *left free to make their own choices*. The game form approach cuts through these problems by representing that principle in a different conceptual framework—a framework that does not refer to individuals' preferences and does not look for social rankings of social states.

4 Latent preferences and the reconciliation problem

I now return to the reconciliation problem. In 2003, two remarkably similar proposals for tackling this problem were published in American legal journals (Camerer et al. 2003; Sunstein and Thaler 2003). Each of these papers had a prominent legal scholar (Cass Sunstein in one case, Samuel Issacharoff in the other) as one of its authors. The other authors were leading American behavioural economists: Richard Thaler (writing with Sunstein) and Colin Camerer, George Loewenstein, Ted O'Donoghue and Matthew Rabin (all writing with Issacharoff). The titles of these papers were similar in intent: 'Libertarian paternalism' and 'Regulation for conservatives'. The common idea was that behavioural economics could justify interventions in the economy (paternalism, regulation) that neoclassical economists had traditionally opposed, and that those justifications were immune to the kinds of objections that might be raised by thinkers on the political right (libertarians, conservatives). Both papers were written as manifestos for an approach that I will call *behavioural welfare economics*. This approach has subsequently become mainstream within behavioural economics and, especially after its popularisation in *Nudge* (Thaler and Sunstein 2008), widely accepted in policy-making circles.

This approach is premised on the assumption that (in Thaler and Sunstein's words) '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'. Public policy, it is argued, should be designed to counter individuals' tendency to make bad decisions. Crucially, its aim should be to 'make choosers better off, *as judged by themselves*' (Thaler and Sunstein 2008, p. 5; italics in original). Or, as Thaler (2015, p. 326) puts it, he and Sunstein 'just want to reduce what people would themselves call errors'. The implication is that an individual's judgements about what makes her better off are expressed in the choices she herself would make in the absence of errors induced by inadequate attention, information, cognitive ability and self-control. By abstracting from the effects of these errors on observable choice behaviour, analysts can reconstruct individuals' *latent* (or 'underlying' or 'true') preferences and use those preferences as data for normative economics. Similarly, Camerer et al. (2003, pp. 1217–1218) say: '[A] large part of behavioural economics describes ways people sometimes fail to behave in their own best interests… It is such errors—apparent violations of rationality—that can justify the need for paternalistic policies to help people make better decisions and come closer to behaving in their own best interest'.

Camerer et al. present behavioural economics as the latest step in a series of advances in economics, each of which has relaxed some assumption that had previously been a standard feature of economic theories. They characterise the 'simplest models in economics' as assuming perfect competition, perfect information and perfect rationality. From the 1930s, economists began to develop models of imperfect competition. From the 1970s, economists began to develop models of imperfect information. Now (as viewed from 2003), behavioural economics is making a further advance: 'Relaxing the assumptions of perfect *rationality* represents a logical next step in this productive progression. The scientific consolidation of psychological findings into a new brand of behavioural economic theory breathes new life into the rationales for paternalistic regulation' (p. 1218, italics in original).

Although Camerer et al. do not quite say this explicitly, the suggestion seems to be that the assumptions of the pre-1930 models supported the view that the market system, if left alone, would tend to produce socially desirable results, and that governments should therefore take a broadly *laissez faire* approach to the regulation of markets. Models of imperfect competition and imperfect information in turn provided rationales for new and more interventionist forms of regulation. Camerer et al. are presenting behavioural economics not only as a further stage in the progress of scientific understanding, but also as a further stage in the recognition of sources of market failure and of the need for interventionist economic policies. As I sense it, the sub-text—one that is common to a lot of work in behavioural welfare economics—is that the findings of behavioural economics undermine traditional arguments in favour of markets.

How do Sunstein and Thaler's and Camerer et al.'s approaches relate to neoclassical welfare economics? It seems clear that both sets of authors are trying to arrive at normative judgements about what is better or worse for society, made from a neutral viewpoint. In Sunstein and Thaler's (2003) paper, this is the viewpoint of a 'planner'—someone who 'must design plans for others' (Sunstein and Thaler 2003, p. 1190); in *Nudge*, it is the viewpoint of a 'choice architect'. Camerer et al. do not say explicitly *who* (apart from themselves) ought to make these judgements but, from the first sentence onwards, it is clear that their topic is 'regulation by the state' (p. 1211). By implication, their viewpoint is that of a publicly authorised regulator. The idea of making neutral judgements about the social good fits with one of the standard interpretations of the social rankings that feature in neoclassical welfare economics, as described in Sect. 2.

For Sunstein and Thaler, at least, it is clear that the data from which these social rankings are to be constructed are individuals' subjective judgements about their own welfare.[2] This matches one of the standard interpretations of the individual preferences that are the basis of neoclassical welfare economics. In neoclassical welfare economics, preferences are assumed to be complete and integrated. Whenever Sunstein and Thaler draw concrete recommendations about public policy, they assume that individuals' judgements about their own welfare have those same formal properties. The difference is that, in neoclassical economics, preferences are assumed to be revealed in choice; there is a background presumption that revealed preferences are broadly consistent with conventional rational choice theory. In contrast, the starting point for behavioural economics is a rejection of that presupposition: far from choosing rationally, individuals make pretty bad decisions. Sunstein and Thaler close the gap by postulating *latent* preferences that have the properties that neoclassical economics attributes to *revealed* preferences. The concepts of 'error' and 'bias' are then brought in to explain why latent preferences are not revealed in choice.

However, much of the evidence that Sunstein and Thaler present in support of the claim that individuals make bad decisions is not evidence of identifiable errors of knowledge or reasoning. Some of it is merely evidence that people make decisions that, as judged by those authors (and probably by many of their readers) are imprudent

[2] Camerer et al. are less precise about the data they want to use. In their account, a fully rational individual acts on well-defined preferences which represent her 'best interests'. Her best interests correspond with 'true costs and benefits' to her, but 'cost' and 'benefit' are never defined (pp. 1214–1215).

or foolish. For example, in arguing for nudges against obesity-inducing diets, drinking and smoking, Thaler and Sunstein (2008, pp. 7, 44) simply report familiar statistics about the associated health risks and then conclude: 'With respect to diet, smoking, and drinking, people's current choices cannot reasonably be claimed to be the best means of promoting their well-being' (pp. 7, 44). Notice the unstated assumption that a person's choices *ought* to reveal her judgements about the best means of promoting her well-being. The obese forty-a-day smoker (we are told) cannot reasonably believe that he is choosing what is best for his well-being, and therefore *must* be making what he himself would call an error. (My guess is that Sunstein and Thaler would hypothesise that the error was a failure of self-control, acknowledged as such by the smoker himself. But perhaps the smoker has unreasonable beliefs that seem reasonable enough to him. Or perhaps he is perfectly well aware that he is not promoting his long-term well-being but just doesn't feel any inclination to change his behaviour.)

Other evidence, which Sunstein and Thaler interpret as showing the effect of 'biases', is more accurately characterised as evidence that individuals' choices are influenced by contextual features of the decision-making environment that irrelevant for the choosers' welfare. Consider their favourite example of the cafeteria customer whose choices between food items depend on the prominence with which they are displayed (pp. 1–2). The psychological mechanism at work here is clear enough: when choosing among positively-valued items, people are more likely to choose those to which their attention is most directed. But that does not tell us how much attention a person should give to each item, and so does not tell us which choices are biased and which are not. The idea of bias presupposes a concept of truth, but empirical psychology does not provide—and has no need for—a concept of true preference.

My diagnosis is that Sunstein and Thaler (and behavioural welfare economists more generally) are running up against difficulties that are inherent in their conception of latent preference. If latent preferences are to serve the purposes for which they have been invoked, each individual's latent preferences need to have four properties that do not easily coexist. First, they must be complete and integrated (at least, within the policy domains to which they are to be applied). Second, they must represent that individual's subjective judgements about her own well-being. Third, they must represent the choices that that individual would in fact make in the absence of identifiable errors of information, reasoning, attention or self-control. Fourth, those errors must be ones that the individual herself recognises as errors. In *The Community of Advantage*, I say more about the difficulties of using latent preferences as the building blocks of normative economics.[3]

5 How the two problems are similar

I can now explain the parallels that I see between the two sets of problems I have discussed—those that social choice theorists faced in the 1970s and 1980s in formulating principles of individual liberty, and those that behavioural welfare economists now confront when they use the concept of latent preference.

[3] My discussion of this topic in the book is based on Infante et al. (2016).

Ever since Adam Smith's *Wealth of Nations*, there has been a tradition of liberal economics in which competitive markets have been viewed favourably. Something of that intuitive idea is expressed in Smith's metaphor of the invisible hand that leads market participants to promote the interest of society even though that is no part of their intentions (Smith 1776/1976, p. 456). A different aspect of that intuition is expressed in Smith's characterisation of the market as a system of 'natural liberty', in the more modern metaphor of consumer sovereignty, and in economists' long-standing professional aversion to paternalism. Roughly, the thought is that markets provide individuals with opportunities to buy whatever they would like to buy and to sell whatever they would like to sell. In providing these opportunities, markets respect individuals' authority to judge what is best for themselves without having to justify those judgments to anyone else.

Since the 'marginal revolution' of the 1870s, economists have represented these pro-market intuitions in a theoretical framework in which rational agents act on well-defined preferences. *That representation* works only to the extent that neoclassical assumptions about individuals' preferences are seen as an adequate approximation to reality. The findings of behavioural economics cast doubt on the adequacy of those assumptions. Many behavioural welfare economists seem to have concluded that *the intuitive ideas* expressed in the two metaphors are applicable only if preferences have the neoclassical properties. Recall the same slippage between models and the reality they represent in social choice theorists' interpretations of Sen's impossibility result.

There are also similarities between the theoretical strategies that have been used in response to the two problems. Recall how some social choice theorists concluded that an individual's liberties could or should be protected only if his preferences were non-meddlesome. Analogously, some behavioural welfare economists conclude that sovereignty can or should be allowed only to rational consumers. Camerer et al. (2003) draw this conclusion particularly starkly. They propose an approach of 'asymmetric paternalism' in which the fundamental asymmetry is between 'fully rational' consumers, who should be left free to make their own choices, and 'boundedly rational' consumers, who should be the target of paternalistic regulations.

Recall how other social choice theorists concluded that principles of liberty should be defined so that social rankings respect individuals' preferences only after those preferences have been laundered to remove elements of meddlesomeness. Analogously, the dominant view in behavioural welfare economics is that social welfare judgements should respect individuals' preferences only after those preferences have been laundered to remove the effects of error. In both cases, individuals' actual preferences are being found not to fit a pre-existing theoretical scheme. In each case, the diagnosis is not that there is a flaw in that scheme, but that the preferences that fail to fit it are at fault.[4]

The game form approach resolves the problem posed by Sen's impossibility result by moving outside the conceptual framework of preferences and social rankings and

[4] This feature of some common arguments in behavioural economics has been pointed out by Berg and Gigerenzer (2010), Infante et al. (2016) and Rizzo and Whitman (2020).

instead considering relationships between the choices that individuals make and the outcomes that result from those choices. If one wants to know whether liberal intuitions about the desirable properties of markets remain true, given the findings of behavioural economics, it is natural to use a similar approach. That is the approach that I have been following since the early 2000s, and which has led to *The Community of Advantage*. I do not try to attribute true preferences to people who often seem not to have such things. I do not imagine myself as a benevolent social planner or regulator, trying to maximise a neutral conception of the social good. I ask how far markets provide individuals with opportunities for interactions that, as viewed by each participant and for whatever reason he or she thinks important at the time, are mutually beneficial.

Acknowledgements This paper is based on a talk that I gave at an HEIRS (Happiness Economics and Interpersonal Relations) conference in Rome in November 2019. My work has received funding from the Economic and Social Research Council (award ES/P008976/1) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, Grant agreement No. 670103.

References

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

Berg N, Gigerenzer G (2010) As-if behavioral economics: neoclassical economics in disguise? Hist Econ Ideas 18:133–166

Camerer C, Issacharoff S, Loewenstein G, O'Donoghue T, Rabin M (2003) Regulation for conservatives: behavioral economics and the case for 'asymmetric paternalism'. Univ Pa Law Rev 151:1211–1254

Gaertner W, Pattanaik P, Suzumura K (1992) Individual rights revisited. Economica 59:161–177

Gärdenfors P (1981) Rights, games, and social choice. Noûs 15:341–356

Gibbard A (1974) A Pareto-consistent libertarian claim. J Econ Theory 7:338–410

Infante G, Lecouteux G, Sugden R (2016) Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. J Econ Methodol 23:1–25

Jones P, Sugden R (1982) Evaluating choice. Int Rev Law Econ 2:47–65

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–291

Loewenstein G (1999) Experimental economics from the vantage-point of behavioural economics. Econ J 109:F25–F34

Loomes G, Sugden R (1982) Regret theory: an alternative theory of rational choice under uncertainty. Econ J 92:805–824

Loomes G, Sugden R (1986) Disappointment and dynamic consistency in choice under uncertainty. Rev Econ Stud 53:272–282

Mill JS (1871/1909) Principles of political economy, 1st edn. Longmans, London 1848

Nozick R (1974) Anarchy, state, and utopia. Basic Books, New York

Rawls J (1971) A theory of justice. Harvard University Press, Cambridge

Rizzo M, Whitman G (2020) Escaping paternalism: rationality, behavioural economics and public policy. Cambridge University Press, Cambridge

Sen A (1970) The impossibility of a Paretian liberal. J Polit Econ 78:152–157
Sen A (1982) Choice, welfare and measurement. Basil Blackwell, Oxford
Sen A (1985) Commodities and capabilities. North-Holland, Amsterdam
Smith A (1776/1976) The wealth of nations. Oxford University Press, Oxford
Sugden R (1978) Social choice and individual liberty. In: Artis M, Nobay A (eds) Contemporary economic analysis. Croom Helm, London
Sugden R (1985) Liberty, preference and choice. Econ Philos 1:213–229
Sugden R (1998) The metric of opportunity. Econ Philos 14:307–337
Sugden R (2018) The community of advantage: a behavioural economist's defence of the market. Oxford University Press, Oxford
Sunstein C, Thaler R (2003) Libertarian paternalism is not an oxymoron. Univ Chic Law Rev 70:1159–1202
Thaler R (2015) Misbehaving: how economics became behavioural. Allen Lane, London
Thaler R, Sunstein C (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven