Sugden, Robert

Working Paper

The behavioural economist and the social planner: To whom should behavioural welfare economics be addressed?

Papers on Economics and Evolution, No. 1121

Provided in Cooperation with:
Max Planck Institute of Economics

Suggested Citation: Sugden, Robert (2011) : The behavioural economist and the social planner: To whom should behavioural welfare economics be addressed?, Papers on Economics and Evolution, No. 1121, Max Planck Institute of Economics, Jena

This Version is available at:
https://hdl.handle.net/10419/57559

PAPERS on Economics & Evolution

1121

The behavioural economist and the social planner: to whom should behavioural welfare economics be addressed?

by

Robert Sugden

The *Papers on Economics and Evolution* are edited by the Evolutionary Economics Group, MPI Jena. For editorial correspondence, please contact: evopapers@econ.mpg.de

ISSN 1430-4716

Max Planck Institute of Economics
Evolutionary Economics Group
Kahlaische Str. 10
07745 Jena, Germany
Fax: ++49-3641-686868

The behavioural economist and the social planner:

to whom should behavioural welfare economics be addressed?

Robert Sugden

School of Economics
University of East Anglia
Norwich NR4 7TJ
United Kingdom
r.sugden@uea.ac.uk

20 December 2011

Abstract

This working paper is a lightly edited version of two chapters of a book that I am currently writing. This book will present and defend a form of normative economics that conserves the main insights of the liberal tradition of classical and neoclassical economics but does not depend on strong and implausible assumptions about individual rationality. In this paper, I ask who the addressee of normative economics should be. Conventional welfare economics, both neoclassical and behavioural, asks what is good for society from an impartial perspective – the 'view from nowhere'. Explicitly or implicitly, its recommendations are addressed to an imagined benevolent despot. I argue for an alternative, contractarian approach, in which recommendations are addressed to individuals who are looking for ways of coordinating their behaviour to achieve mutual benefit. The contractarian approach disallows paternalistic recommendations, since these have no valid addressee.

For the last seventy-five years, the main tradition of normative economics has been that of neoclassical welfare economics. Welfare economics is in direct line of descent from the utilitarian philosophy espoused by many of the classical and neoclassical economists of the nineteenth century. It aims to evaluate alternative states of affairs for a society from an impartial point of view. It tries to answer the question: 'What is good for society, all things considered?' It takes the position that the good of society is made up of the good or welfare of each of the individuals who comprise that society. Thus, welfare economics has to assess what is good for each person, all things considered, and then aggregate those assessments. How assessments of individual welfare should be aggregated has been one of the core theoretical problems of welfare economics, for which there is still no universally accepted solution; but that problem is orthogonal to the topic of this paper. For many years, however, there was general agreement on the criterion for assessing what is good for each individual, considered separately. The traditional criterion is preference-satisfaction: if some individual prefers one state of affairs to another, the former is deemed to be better for him than the latter.

This consensus has been disturbed by recent developments in experimental and behavioural economics. As usually applied, the criterion of preference-satisfaction presupposes that each individual has well-formed and reasonably stable preferences over the social states that welfare economics needs to assess. By interpreting those assumed preferences as expressing the individual's judgements about what is good for him, welfare economics can provide a reasonably persuasive justification for the preference-satisfaction criterion. But that presupposition has been called into question by the findings of behavioural economics. Those findings suggest that individuals often come to decision problems without well-defined preferences that pre-exist the particular problem they face; instead, whatever preferences they need to deal with that problem are constructed in the course of thinking about it. Such 'constructed' preferences can be influenced by features of the framing of the problem that seem to have no bearing on the individual's well-being. As a result, the preferences that an individual reveals with respect to given objects of choice (for example, preferences over given bundles of consumption goods) can vary across decision problems according to apparently arbitrary differences of framing. Often, the influence of framing can be explained by reference to the decision-making heuristics that the individual uses to process different decisions problems. But however valuable those heuristics may be

in helping an individual with limited cognitive powers to navigate a complex world, it is difficult to maintain that the preferences they construct are the individual's considered judgements about his welfare.

Given the underlying logic of welfare economics, a natural response to this problem is to supplement the preference-satisfaction criterion with some other principle for assessing individual welfare, applicable where individuals lack well-formed preferences. To remain as faithful as possible to the spirit of traditional welfare economics, one might try to find some way of inferring or reconstructing an individual's underlying judgements about what is good for him from whatever evidence seems most relevant. This, in broad-brush terms, is the approach that most behavioural economists seem to favour. In different variants, it has been called *libertarian paternalism* (Sunstein and Thaler, 2003a, 2003b; Thaler and Sunstein, 2008), *asymmetric paternalism* (Camerer et al., 2003; Loewenstein and Ubel, 2008), and *behavioural welfare economics* (Bernheim and Rangel, 2007, 2009); the general approach is coming to be called *soft paternalism*.

I have proposed an alternative strategy for reconciling behavioural and normative economics (Sugden, 2004b, 2008, 2010; McQuillin and Sugden, 2011). One fundamental respect in which this proposal differs from soft paternalism is that it uses opportunity rather than preference satisfaction as its normative criterion. But there is another, perhaps even more fundamental difference: it has a different addressee. In this paper, I explain and defend this feature of my proposal.

1. The view from nowhere and the benevolent despot

Soft paternalism and neoclassical welfare economics have an important feature in common – the *viewpoint* from which assessments of welfare are made. Because welfare economists are so used to imagining themselves occupying this viewpoint, they tend not to notice just how peculiar it is.

What is peculiar about it? The first thing to notice is that the viewpoint is *synoptic*: it is the viewpoint of a single viewer, who is not any of the individual people who comprise the society that is being assessed. The viewer somehow stands outside society and makes judgements about its overall goodness. This is the kind of view that has traditionally been attributed to God, looking down on his creation. To use a phrase coined by Thomas Nagel (1986), it is a 'view from nowhere'. (What else can it be, if it is to encompass everything?)

Nagel thinks that this is exactly the viewpoint that we *should* take when we try to engage in moral reasoning. The thought is that, when a person thinks morally, he somehow rises above his ordinary self and assumes a viewpoint from which he can see that self as just one person among others. But I cannot resist borrowing Nagel's words and giving them a sceptical intonation. A view from nowhere is, to put it mildly, a peculiar view.

The welfare economist's viewpoint, then, is that of a *spectator* – someone who views society from outside. Since the point of taking this viewpoint is to try to filter out one's private interests and biases, it is crucial that the imagined spectator is *impartial* with respect to the preferences and interests of the various individuals whose welfare she is assessing. And since the aim is to assess welfare, the spectator must be assumed to take an interest in the welfare of every individual who comes into her synoptic view. So the welfare economist has to imagine an *impartially benevolent spectator*.[1]

Suppose we accept the meaningfulness of the view from nowhere. Suppose we have found a method of assessing the good of society, all things considered, as viewed by an impartially benevolent spectator. What then? Who is supposed to use this assessment, and for what purpose?

One possible answer, sometimes proposed by utilitarian moral philosophers, is that every individual ought to act with the objective of maximising the overall good of society (or, better, the overall good of the universe). I have to say that this is not an idea that appeals to me. My internal sense of morality is of particular obligations and commitments that arise out of particular relationships between me and the rest of the world. I do not feel an unconditional obligation to give just as much weight to anyone's interests as I give to my own, or to those of my own family, friends and colleagues; and nor do I expect unrelated others to feel such obligations to me. But perhaps that just reveals my moral limitations. In any case, there is no need to pursue this line of thought. What is at issue here is what *welfare economists* do with their assessments of the social good. Welfare economics, as it is normally practised, is not about the moral obligations of private individuals.

[1] Some readers may think, as John Rawls (1971: 184–185, 263–264) seems to do, that this conception of the impartially benevolent spectator is the one used by Adam Smith in his *Theory of Moral Sentiments* (1759/ 1976). But Smith's impartial spectator is a representative human being, whose sympathies for other people are governed by the mechanisms of real human psychology and so incorporate those mechanisms' natural biases. Smith is not taking a view from nowhere; he is proposing a naturalistic theory of the moral sentiments that people in fact feel.

The traditional addressee of welfare economics is an entity variously known as ‘the ‘the policy-maker’, ‘the government’ or ‘the social planner’. (An outsider might be surprised that social planners still have their place in the dramatis personae of economics, but they do.) In an alternative formulation of the same basic idea, applied economists often end their papers by drawing ‘policy implications’ from their analyses, these being the actions that the policy-maker is recommended to take. The implicit assumption is that this addressee is, or ought to be, motivated by concern for the overall good of society, as viewed by an impartially benevolent spectator.

This understanding of the purpose of normative economics has been carried over to behavioural welfare economics in its various guises. Thus, in their first presentations of libertarian paternalism, Cass Sunstein and Richard Thaler conceive of themselves as addressing a ‘planner’, defined as ‘anyone who must design plans for others, from human resource directors to bureaucrats to kings’ (2003a: 1190). More recently, perhaps recognising the negative connotations of social planning, they have renamed their addressee as a ‘choice architect’, but the job specification remains the same (Thaler and Sunstein, 2008). They focus on the role of the choice architect in designing the formats in which decision problems are presented to individuals. If, as the behavioural evidence suggests is often the case, individuals’ choices are sensitive to variations in decision formats, Sunstein and Thaler’s addressee has the power to influence what individuals choose. How should she use this power?

Using the example of a cafeteria director deciding how to display different food items, knowing that different displays will induce different choices on the part of her customers, Sunstein and Thaler (2003a: 1164) interpret traditional welfare economics as recommending that she should ‘give consumers what she thinks they would choose on their own’. (Notice how the concept of *giving* is being used here: I will come back to this.) But this recommendation cannot help the cafeteria director, because what the customers will choose ‘on their own’ can be defined only relative to the decision format, and the whole problem is to decide what this format should be. Sunstein and Thaler conclude that the director should choose the format that ‘she thinks would make the customers best off, all things considered’, subject to the constraint that freedom of choice is not restricted. By virtue of this constraint, Sunstein and Thaler’s recommendation ensures that individuals get what they prefer whenever their preferences are independent of the decision format. Thus, one might say, libertarian paternalism agrees with traditional welfare economics whenever

the well-formed preferences assumed by the latter exist; when they do not, libertarian paternalism uses a well-being criterion that is consistent with the spirit of traditional welfare economics. The close relationship between the two forms of welfare economics reflects their common conception of normative economics as addressed to an impartially benevolent social planner.

Douglas Bernheim and Antonio Rangel's (2007, 2009) behavioural welfare economics follows a similar logic. Bernheim and Rangel are explicit in invoking a planner. They interpret 'standard welfare analysis' as 'instruct[ing] the planner to respect the choices an individual would make for himself'. This normative principle is presented as 'an extension of the libertarian deference to freedom of choice, which takes the view that it is better to give a person the thing he would choose for himself rather than something that someone else would choose for him' (2007: 464). (Notice again the idea that individuals' freedom of choice can be represented in terms of what a planner *gives* them.) Like Sunstein and Thaler, Bernheim and Rangel see the findings of behavioural economics as revealing ambiguities in the concept of what a person would choose for himself. If an individual's behaviour shows a 'choice reversal' – that is, if she would choose object *x* over object *y* under some conditions, but *y* over *x* in others – then her choices 'fail to provide clear guidance' to the planner (p. 465). What is required in such cases, therefore, is some set of criteria 'to officiate between conflicting choice data' (p. 469); one of the aims of behavioural welfare economics is to provide such criteria.

So welfare economics, in both its traditional and behavioural forms, is addressed to an imagined policy-maker. The presumption must be that this policy-maker will find some use for the welfare economics that is addressed to her. But what use?

As James Buchanan has often said (and has attributed to the earlier writings of Knut Wicksell), welfare economics is implicitly addressed to a benevolent despot (e.g. Buchanan, 1986: 23). The imagined policy-maker must be impartially benevolent if she is to have the motivation to act on the policy implications she is being informed about. In her public role, she must treat the social good, impartially assessed, as her only objective. She must give no weight to her private career interests, or (if she is an elected politician) to her chances of being re-elected. But impartial benevolence is not enough. If she is to be able to implement whatever policies maximise the overall good of society, we must imagine her to have the powers of an enlightened despot. We must imagine that she is not subject to the messy constraints that political leaders and civil servants have to face in real-world democracies.

Having recognised that a certain policy is the best, she does not have to negotiate with other members of her cabinet or party who might disagree with her. She does not have to take the policy to a Parliament or Congress where it might be voted down. She simply gives the order that the policy is to be implemented, and moves on to the next problem in her in-tray.

There is a further sense in which the imagined policy-maker is unconstrained. Recall how, both for Sunstein and Thaler and for Bernheim and Pearce, the idea of respecting individuals' preferences is represented in terms of the policy-maker *giving* individuals what they prefer. This is not a wholly innocent figure of speech. The social planner to whom welfare economics is addressed is not supposed to be *constrained by* individuals' preferences. She may choose to *take account of* those preferences, and welfare economics advises her on how to do so; but whether she acts on this advice is up to her. And so whether individuals get what they prefer depends on how the planner uses her discretionary power. If they do get what they prefer, that is as a result of the planner's decisions, for which she takes responsibility. In this sense, she is deciding what individuals are to be given: they are not deciding for themselves what they are to have.

There is yet more to the fiction. Even if the imagined policy-maker were impartially benevolent and had the powers of an enlightened despot, she might still not want to act on the welfare economist's recommendations. Take the example of the cafeteria again. In this case, Sunstein and Thaler are playing the role of the welfare economist, advising on the display of food items; the cafeteria director is the addressee of their advice. The problem, as Sunstein and Thaler formulate it, is to choose the display that maximises the welfare of the cafeteria customers, all things considered. Solving the problem involves making contestable judgements. To start with, there is no uniquely correct concept of welfare. In assessing people's welfare, Sunstein and Thaler seem to want to use what philosophers call an 'informed desire' criterion – that is, they want to assess welfare by reference to what people would choose if they had 'complete information, unlimited cognitive abilities, and no lack of willpower' (2003a: 1162). Already, Sunstein and Thaler are taking a philosophical position that the policy-maker might not share. (She might favour a different conception of impartial benevolence, such as the maximisation of happiness.) To specify what a person would choose in the light of 'complete information', one has to make scientific judgements about the best inferences to draw from the available evidence. In the cafeteria problem, judgements have to be made about how variations in diet affect health and life expectancy. On this issue, different scientists make different judgements. A welfare economist who is

confident that one dietary theory is correct may find himself advising a policy-maker who is equally confident about a different theory. And so on.

When welfare economists talk about 'policy implications', they normally use *their own* best judgements about contestable normative and scientific questions. Unless they are working as paid consultants (in which case they are addressing real policy-makers, not imagined ones), they do not ask whether these judgements are shared by their addressees. The implicit thought is that if the welfare economist uses *his own* best judgements, he is entitled to assume that the policy-maker will accept these as *the* best judgements. So the imagined policy-maker is not just an impartially benevolent despot: she is an impartially benevolent despot who, on all contestable normative and scientific questions, agrees with the welfare economist who is advising her. But if this is so, the conceptual distinction between adviser and policy-maker evaporates. We might as well say that the welfare economist is imagining that *he* is the benevolent despot. The content of a policy implication is: *If I were an impartially benevolent despot, this is what I would do.*

Of course, welfare economists do not *really* believe that their work is being read by an impartially benevolent despot who thinks as they do on all controversial questions and is eagerly waiting for their advice. Nor, typically, do they think of benevolent despotism as an ideal political system, to which actual procedures of collective choice are imperfect approximations. Their recommendations are *not intended to be taken literally*.

Suppose that, in my capacity as a welfare economist, I have been commissioned to write a report for a government department, advising on some issue of economic policy. My report recommends some course of action – say, the compulsory metering of domestic water supplies – which makes good economic sense to me but to which, for what I believe to be mistaken reasons, many people object. The politician who heads the department tells me that she agrees with my analysis, but judges my proposal too unpopular to implement. In other words, if she were an impartially benevolent despot, she would act on my advice; but she is not. That does not make my advice mistaken or useless: we might both think that it is useful to look at the problem from the perspective of conventional welfare economics, while recognising that this is not the only perspective that is relevant for a democratic politician. But notice that I am not advising her to ignore the political constraints to which she is subject. I am not advising her to investigate the feasibility of coup. I am not suggesting that she should commission me to report on whether a seizure of power would increase social welfare, all things considered. In the literal sense, I am not advising her to implement the

policy I am 'recommending'. I am merely telling her that this is a recommendation that I would act on, were I an impartially benevolent despot.

So the idea of the impartially benevolent despot as the addressee of welfare economics is not an assumption about the powers of any real person or institution. It is a framework for organising thought, a literary device. In the language of economics, it is a *model*.

For the present, that is all I need to say. It is sufficient to recognise that the impartially benevolent despot belongs in a model world, and that all model worlds are unrealistic. To understand the argument I will develop later in this paper, the reader must be able to step outside the traditional model and see that it is not the only way of thinking about normative economics. I will present an alternative model in which there is a different addressee (or as will become clear, addressees). I will ask the reader to consider the two models side by side, and not to criticise my approach on the grounds that it fails to give the right recommendations to the impartially benevolent despot that the traditional model imagines. Of course it does: it is not addressed to her.

2. Public reasoning

I began the previous section by asking what the point of the view from nowhere was supposed to be. What was the use of impartial assessments of the good of society, all things considered? I argued that, in both traditional and behavioural welfare economics, such assessments are construed as recommendations to an imaginary benevolent despot. However, this is only one way of using the idea that normative reasoning requires a view from nowhere. Since I want to persuade readers to set aside this fundamental presupposition, I need to consider other ways in which normative economics might be grounded on a view from nowhere. To keep the discussion concrete, I will focus on the work of one of the most influential critics of traditional welfare economics, Amartya Sen.

As a starting point, I take the 'parable' with which Sen (1999: 54–58) introduces a wide-ranging analysis of freedom and justice. The story is of a woman hiring a labourer to work in her garden. There are three applicants, all currently unemployed, and each of whom would do much the same work for the same payment. '[B]eing a reflective person', the employer 'wonders who would be the right person to employ'. Sen imagines the employer asking herself how, in choosing between the applicants, she can do the most good. Should

she choose Dinu, the poorest applicant (thus doing as much as she can to reduce poverty)? Or should she choose Bishanno, the applicant who would gain most happiness from being employed (thus doing as much as she can to increase happiness)? Or should she choose Rogini, the applicant for whom the job would make the biggest difference to 'the quality of life and freedom from illness'? The purpose of the story is to set out three apparently credible 'evaluative approaches' to normative economics, each of which has a different 'informational basis'. If the only available information was about income, there would be a good 'income-egalitarian case' for hiring Dinu. If the only available information was about happiness, there would be a good 'classical utilitarian case' for hiring Bishanno. And if the only available information was about health-related deprivation, there would be a good 'quality-of-life case' for hiring Rogini. Sen sees each of these cases as having its merits, and tries to find a normative framework that can encompass their different informational bases.

Clearly, Sen's approach to normative analysis is much wider than that of conventional welfare economics, which he sees as having a particularly impoverished informational basis. But it is still a view from nowhere. Sen's story is about alternative ways of distributing a valuable resource between three needy individuals; the suggestion is that this is a miniature version of one of the central problems of normative economics. Significantly, he presents this problem through the eyes of a fourth person who, from a neutral position, 'reflectively' asks which solution would be best – not best from her private viewpoint as an employer, but best in some impartial sense.

Thus (Sen tells us), when the employer thinks about Dinu, she asks herself: 'What can be more important than helping the poorest?' Similarly, when she thinks about Bishanno, she tells herself: 'Surely removing unhappiness has to be the first priority'. One might ask: important *for whom*? First priority *for whom*? I take it that, for Sen, these questions would be superfluous. He is not talking about what is important *for* anyone in particular; he is talking about what just *is* important. This is a view from nowhere, the view as seen by some kind of impartial spectator.

It would perhaps be wrong to describe Sen's imagined spectator as impartially *benevolent*, since that might suggest that she takes the evaluative approach of classical utilitarianism, and Sen sees utilitarianism only as one eligible approach among others. But we must imagine the spectator to be sympathetically or morally engaged with the society on which she is looking, while not being part of it. She is concerned that the state of this society should be good rather than bad. From her impartial viewpoint, she recognises that

income equality, happiness, quality of life and freedom all contribute to the good of society all things considered. Her problem is to reach an impartial assessment of the relative importance of these different forms of goodness.

What, then, is the point of arriving at an impartial assessment of overall goodness? Who is supposed to use it, and for what purpose?

Unlike traditional and behavioural welfare economists, Sen does not imagine himself addressing a social planner. He addresses individuals *as citizens*, participating in public discussion – or, as he often says, in public reasoning – about the social good. In democratic societies, such discussions may influence and perhaps even determine collective choices, but Sen wants to be able to contribute to public reasoning about *any* society, democratic or undemocratic. Indeed, it is particularly important for Sen that the kind of normative discourse in which he is engaging can be used to diagnose injustice anywhere in the world, and so can be used to 'fight oppression ..., or protest against systematic medical neglect ..., or repudiate the permissibility of torture ..., or reject the quiet acceptance of chronic hunger' (2009: xi – xii). Some of the alleged injustices on Sen's charge sheet (for example, the deficiencies of health care provision in the United States, and the persistence of hunger in India) are practised in open and democratic societies, but when he attacks injustices committed by authoritarian regimes, the public discussion to which he is contributing is presumably one of opposition: it is certainly not the policy-making process.

For Sen (2009: 39–46), public reasoning involves debate about 'the demands of ethical objectivity'. The implication is that there can be objectivity in ethics, and that objectivity makes 'demands' on us as citizens which in some sense we are required to meet. But what does this mean?

On the most natural reading, objectivity in ethics requires that there are ethical *objects*, and that what these objects consist of is a matter of fact. Some important philosophical traditions do maintain exactly this, seeing moral truths as somehow part of the fabric of the universe. In some traditions of natural religion, moral truths are woven into that fabric because God put them there; we can discover them by inferring God's benevolent purposes from the well-designedness of the universe and then considering what is necessary for those purposes to be achieved. In much Enlightenment thought, the will of God is replaced by reason (or even 'Reason' with a capital 'R'). The idea is that rational beings (a category to which *homo sapiens* is of course supposed to belong) have a common faculty

which allows each of them to recognise those propositions that are implications of reason. Moral truths are supposed to be accessible in this way; since the content of reason is supposed to be the same for all rational beings, moral truths are objective.

The difficulty with these ways of thinking about ethics is that they postulate the existence of moral or rational objects whose properties are quite unlike those that empirical science recognises, while providing us with no adequate way of verifying whether those objects exist or not. One of the classic statement of such doubts is by David Hume (1739-40/ 1978); a famous and more recent objection to the 'queerness' of objective morality is made by John Mackie in a book with the provocative title *Ethics: Inventing Right and Wrong*. As Mackie puts it: 'If there were objective values, then they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe' (1977, p. 38). I share his scepticism.

A recurring theme in recent moral philosophy is the attempt to find some way of understanding ethical objectivity that does not presuppose the existence of ethical objects. Drawing on the work of Smith (1759/ 1976), John Rawls (1993), Jürgen Habermas (1995) and Hillary Putnam (2004), Sen locates objectivity *in* public reasoning. By this I mean that, for Sen, public reasoning is not to be understood as an attempt to *discover* objective truths about ethics that have an independent existence. Rather, ethical propositions are objective *by virtue of their being certified by* the right kind of public reasoning. Sen argues that, in the work of Smith, Rawls and Harbermas, 'objectivity is linked, directly or indirectly, to the ability to survive challenges from informed scrutiny coming from diverse quarters'. Similarly, Sen says of his own analysis of justice: 'I will take reasoned scrutiny from different perspectives to be an essential part of the demands of objectivity for ethical and political convictions' (2009: 45). How, one might ask, is *reasoned* scrutiny differentiated from *unreasoned*? Sen is not particularly explicit about the standards of reasoning he is invoking, but he is clear that these must include impartiality: 'The reasoning that is sought in analysing the requirements of justice will incorporate some basic demands of impartiality, which are integral parts of the idea of justice and injustice' (2009: 42).

So, for Sen, normative analysis is a contribution to a process of public reasoning in which citizens try to reach agreement on an impartial assessment of the social good. Sen's language of *demands* and *requirements* (the 'demands of objectivity' for ethical and political convictions, the 'requirements of justice', the 'demands of impartiality') seems to imply that engagement in public reasoning is not to be thought of as an optional activity, like joining a

reading group or debating society. Rather, there is some kind of moral requirement on each of us, as citizens or as rational agents, to make impartial assessments of the social good, to defend these assessments by reasoned argument, and to expose these arguments to other people's scrutiny.

As a model of this conception of public reasoning, consider a jury which has to determine an issue of judgement when there is no dispute about facts. (Suppose the defendant has been charged with murder, having fatally shot an intruder to his house; he admits the killing but pleads self-defence. Given the facts of the case, the jury has to decide whether the defendant's use of force was reasonable.) For each member of the jury, participation in the judgement process is a duty of citizenship. She is expected to set aside her private interests, preferences and prejudices and try to reach an impartial judgement about whether the defendant's action was reasonable. Collectively, the jury is expected to engage in reasoned discussion; each member is expected to take account of the others' arguments, while being individually responsible for his or her final decision. Because the twelve members of the jury have been selected at random, there is an expectation that the discussion will be informed by the varied experiences and insights of the members. There is a hope, but not a requirement, that this process will end in agreement.

Sen's account of public reasoning belongs to a tradition of political philosophy in which politics is interpreted on the model of the jury.[2] I do not want to claim that there is anything incoherent in this conception of politics. But it does have some troubling features.

Recall that the objective is to arrive at an impartial assessment of the social good, all things considered, and that the building blocks for this are impartial assessments of the good of each individual. So consider a specific individual: me. Suppose that what is at issue is some decision that I am about to make about how to live my own life – perhaps about how to balance work and leisure, or how to weight immediate enjoyment against future health in choices about eating and drinking. I might acknowledge that, *were I trying to make an impartial assessment of what is good for me*, other people's judgements would be relevant. But why do I need an impartial assessment of what is good for me? What matters *to me* is

[2] The idea that many political philosophers treat politics as a 'generalization of the jury' is a recurring theme in the work of Buchanan (e.g. 1986: 65). Sen (2009: 110–111) reads Buchanan (1954, 1986) as *approving* this conception of politics, but I think this is a misunderstanding. Buchanan has consistently advocated a contractarian conception of politics as a 'generalization of the market' (1986: 65) and has opposed the idea of political discourse as a search for truth as a 'Platonic faith' held by writers who 'play at being God' (1975: 1–2).

my own assessment. In arriving at that assessment, it would perhaps be wise for me to listen to what other people have to say about what they think is good for me; but ultimately how I live my life is my business and not theirs. I am entitled to treat my own judgements as authoritative, not because I believe they would be endorsed by an impartial spectator, but because I am the author of my own life.

Suppose some moral philosopher tells me that I am morally required to justify my private choices by reasoned arguments. He tells me that I am required to defend these choices as good for me in some objective sense, and to expose those defences to public scrutiny. My reply would be: 'But why?' I cannot see where such a requirement can come from. If he tells me it is a demand of objectivity, and that what he means by the 'objectivity' of a proposition is its ability to survive reasoned scrutiny in public debate, I can reply that, when I take decisions about my own life, I am not interested in that kind of objectivity. Subjectivity is good enough for me. In saying that, I am *not* claiming that, from an impartial point of view and all things considered, it is good that each individual is free to make decisions about his own life, and hence that it is good that I am free to make decisions about my life. Were I to make such a claim, the philosopher might perhaps be entitled to expect me to defend it by reasoned argument. But I am not pretending to report any view from nowhere. What I am saying is much simpler than that: all I am saying is that my own view is what matters to me.

A similar argument can be made about the internal affairs of a political community or voluntary association. Sen's analysis of public reasoning emphasises the importance of taking account of outsiders' judgements in arriving at impartial assessments of what is good in any particular society. He proposes a principle of *open impartiality*: 'Impartial views may come from far or from within a community, or a nation, or a culture' (2009: 123). Thus, he urges Americans to listen to public debate in Europe about whether capital punishment is justified (2009: 407). Given that Sen is trying to find a view from nowhere, this stress on open impartiality is entirely natural. If we are asking whether an impartial spectator would approve or disapprove of capital punishment, it is surely relevant to consider judgements made by the inhabitants of jurisdictions with and without the death penalty. But one might still ask whether political debate should be understood as an attempt to achieve the viewpoint of an impartial spectator. When a decision has to be made about the American criminal justice system, shouldn't the judgements that ultimately count be the judgements of Americans – not because their judgements are more likely to be right, but because it is *their*

system? If the members of some community can agree among themselves on how to organise their internal affairs, why do they need to ask whether their decisions would meet the approval of an impartial spectator?

The concept of public reasoning, like that of the benevolent despot, provides a framework within which ideas about normative economics can be organised. Common to both frameworks is the attempt to find a view from nowhere. It is now time to consider whether there is a viable alternative to this approach. I will argue that there is such an alternative: contractarianism.

3. The contractarian perspective

In the sense in which I will use the term 'contractarian', the most fundamental characteristic of this perspective is that recommendations are addressed to individuals, showing them how they can coordinate their behaviour to achieve mutual benefit. In making some recommendation *R* to some set of individuals, the contractarian says: 'It is in the interests of each of you separately that all of you together agree to do *R*'.

Notice that this is *not* the same thing as saying: '*R* is in the collective interests of the group of which you are the members'. The latter recommendation treats the addressees as a collective, and allows the possibility that *R* requires some individuals to incur losses for the greater good of others. In contrast, the contractarian recommendation is about the good of *each*, not about the good of the *whole*. But notice too that the contractarian recommendation aims at *mutual* benefit, and it is about the terms on which individuals should *agree*. For these reasons, it is not just a collection of separate recommendations addressed to separate individuals. It is a recommendation (in the singular) addressed to individuals (in the plural). Although those individuals are not addressed as components of a collective entity, they are addressed *together*.

The stance taken by a contractarian is similar to that of a mediator, helping the parties to a conflict to find a resolution that they can recognise as mutually beneficial. Pursuing this analogy, the stance of the mediator can be contrasted with that of someone who advises one of the parties to a negotiation on how best to achieve his interests, given the likely behaviour of the others. Such an adviser can look for ways in which the party she is advising can out-think the others. Since the contractarian mediator is advising all the parties together, the idea that one might out-think another can have no place in her reasoning. If there is a range of

alternative terms of agreement, all of which ensure positive benefits to all parties but some of which particularly favour one party, some another, a contractarian mediator must appeal to some principle, whether of rationality or fairness or salience, which all parties acknowledge. Hobbes's second law of nature, requiring each man to be contented with as much liberty against other men as he would allow other men against himself, is an example of this kind of contractarian reasoning.

Since contractarian reasoning is about the achievement of mutual benefit through agreement, it necessarily presupposes some *baseline* of non-agreement from which benefit is measured. And since this reasoning is addressed to individuals together, and is intended to engage with each individual's own interests as he perceives them, this baseline must be acknowledged by each individual. That is, each must recognise that all of them together are looking for an agreement that, for each of them separately, will be more beneficial than non-agreement.

Contractarian writers differ on what is involved in this acknowledgement of a baseline. I share the view of James Buchanan (1975) that, for contractarian reasoning to be possible, it is sufficient that individuals acknowledge the baseline *as a fact of life* – that, as Buchanan puts it, 'we start from here, and not from some place else'(p. 78). In Buchanan's theory of 'ordered anarchy', there is a 'natural distribution' of resources that has emerged in a Hobbesian state of nature, as an equilibrium between individuals whose relationships with one another are those of predator and prey. As an example of this kind of baseline, consider the leaders of the two opposing sides in a civil war, trying to negotiate a political settlement after the war has reached a stalemate. Each may believe his own party to be the legitimate government of the country, and entirely deny the moral legitimacy of the other's claims. Still, if each recognises the reality of the stalemate – that warfare is costly for both sides and that neither has a realistic prospect of outright victory – there may be sufficient basis for negotiation, and hence for contractarian reasoning about mutual benefit.

As a less dramatic example of the same idea, consider two private individuals *A* and *B* in a society with reasonably secure property rights, negotiating over the sale of a car; *A* is the potential seller and *B* the potential buyer. If this is a normal market transaction, their negotiation is structured by their common acknowledgement of their existing property rights in the goods – *A*'s car and *B*'s money – that are to be exchanged. This does not mean that each person has to believe that those rights are legitimated by some comprehensive theory of social justice, but only that issues of social justice are bracketed out of their reasoning about

the terms on which they might trade. Thus, whatever the relative wealth of *A* and *B*, and whatever their respective political opinions about how wealth ought be distributed, neither of them expects to trade on terms that impose a net loss on one party for the benefit of the other.

I have claimed that, for contractarian reasoning to be possible, it is sufficient that the parties acknowledge some non-agreement baseline as a fact of life. Nevertheless, one might expect the parties' perceptions of the moral status of their agreement to be influenced by their perceptions of the moral status of the baseline. This is not quite as obviously true as it appears at first sight. A recurring theme in the work of David Hume (1739-40/ 1978), developed in my own contractarian theorising (Sugden, 2004a), is that ongoing conventions can come to be perceived as having moral status, without reference to any beliefs about the fairness of their actual or hypothetical origins. But if, as some contractarian thinkers do, one wants to provide moral justifications for social rules or institutions construed as mutually beneficial agreements, it is natural to make that justification conditional on the fairness of the baseline. For example, David Gauthier (1986), whose contractarian project is to derive 'morals by agreement', requires that the relevant agreements are made from a baseline in which there is no coercion. Even Thomas Hobbes (1651/ 1962: 98–102) claims that his state of nature, which might seem completely devoid of morality, is a state of approximate equality – equality, that is, in the faculties of body and mind that can be used for self-preservation.

For my purposes, it is important to distinguish between a contractarian conception of a fair baseline for agreement and the 'veil of ignorance' constructions used by John Harsanyi (1955) and John Rawls (1971). In different ways, Harsanyi and Rawls create imaginary 'original positions' in which the individuals who are to become the members of a real society do not know any of the facts that, in reality, differentiate them. In such a model, none of the agents knows his or her own abilities, preferences or moral values. Nor do these agents have different beliefs about how the world works; to the extent that the model requires them to have such beliefs, their beliefs are supposed to correspond with what is really true (or with what the modeller takes to be really true). The modeller then attributes rationality to the agents and investigates the choices they would make between alternative properties of the real society that they will join when the veil of ignorance is lifted. The effect of these constructions is to make each agent's relationship to the real society that of an

impartially benevolent spectator. This is the view from nowhere in another guise; it is not contractarianism in the sense that I am using the term.[3]

Another significant feature of contractarian reasoning is that it typically leads to recommendations in favour of *general rules*. When a particular rule is recommended to individuals, the claim is not that each individual benefits from *every* application of that rule, considered separately, but rather that each can expect to benefit *overall* from the general application of the rule – or, as Hume (1739-40/ 1978: 497) puts it, that each can expect to find himself in credit when he balances his account. As a modern example, consider the rule that requires vehicles entering a roundabout to give way to vehicles that have already entered. It is easy to see that this rule is efficient in ensuring smooth traffic flows. (If the opposite rule is used, as is apparently the case in Uzbekistan,[4] there seems to be no way of unravelling a traffic jam at a roundabout, once it has formed.) Nevertheless, if one considers the application of this rule to a specific interaction between two drivers at a particular moment, it benefits one at the expense of the other. A traffic engineer who takes the viewpoint of a social planner might point out that, on average, the gain in time to the driver who is favoured by the rule is greater than the loss of time to the one who is disfavoured, and so recommend the rule as a means of reducing the *total* time spent by all road users making a given set of journeys. Viewed in the contractarian perspective, this is not an adequate recommendation. A recommendation has to be addressed to each individual separately, and each individual's interest is in her own journey times, not in the total. The contractarian argument for the rule is that, because each individual can expect to be favoured by the rule approximately as often as she is not, everyone can expect to benefit.

In the case of the roundabout rule, the formula 'everyone can expect to benefit' can be read as 'if the rule is applied, everyone *will* benefit in the long run'. But there is another way in which contractarian arguments can use the concept of expectation. Consider the rule (whether legal or moral) that if there has been a serious road accident, the first person to arrive on the scene must provide assistance, at least to the extent of calling the emergency

[3] This is a comment on Rawls's model of the original position, and not on his theory of justice as a whole. Some of the core ideas of that theory, in particular that society should be understood as a cooperative venture for mutual advantage, are unquestionably contractarian. My view, for what it is worth, is that Rawls's attempt to combine a morality of mutual advantage with a view from nowhere creates tensions that he ultimately fails to resolve.

[4] At the time of writing, the British government's Foreign and Commonwealth Office website includes the following advice to travellers to Uzbekistan: ' Be aware that vehicles approaching a roundabout have the right of way over vehicles already on the roundabout'. Thanks to my son Joe for directing me to this intriguing information.

services if the accident victims are incapable of doing so. In each specific case, this rule imposes significant costs on the person who has to provide assistance, but provides much greater benefits to the people who are being assisted. Because serious accidents are rare events, it would not be true to say that, for each individual separately, the *ex post* benefits of the rule exceed the *ex post* costs. But one might reasonably claim that, for each individual *ex ante* – that is, looking ahead and considering the probabilities of his being involved in an accident in the two possible roles – the *prospective* benefits exceed the *prospective* costs. To use an idea developed by James Buchanan and Gordon Tullock (1962, pp. 77–81), when an individual considers the future application of a general rule, a *veil of uncertainty* limits his ability to see exactly how that rule will affect him. The veil of uncertainty represents the uncertainty faced by real individuals when thinking about the possible consequences to them of general rules; the view from behind it, unlike the view from behind Rawls's veil of ignorance, is not a view from nowhere.[5] Both because of this uncertainty and for the reason that Hume describes in terms of 'balancing the account', individuals' interests tend to be more closely aligned with respect to general rules than with respect to particular cases.

At first sight, it might seem that the contractarian approach can work *only* when applied to very general rules. If there is to be a contractarian recommendation in favour of a specific policy, it must be addressed separately to every individual who is affected by that policy. How often, a sceptic might ask, do we find policies that benefit some individuals without harming *anyone*?

As a starting point for a response to this kind of scepticism, consider the workings of markets for private goods, as in my example of *A* and *B* negotiating over the sale of a car. In a typical case, such a trade affects only the two parties involved. If *A* and *B* agree to trade at a particular price, it is reasonable to presume that the resulting transaction is beneficial to each of them in terms of her own interests, as she perceives them, and that no one else's interests are materially affected; the relevant benchmark is the allocation of resources prior to trade. By extension, any combination of voluntary exchanges of private goods between

[5] The term 'veil of uncertainty' is not used explicitly by Buchanan and Tullock (1962), but Buchanan uses it in later work when referring to this idea, to draw attention to analogies and disanalogies with Rawls' veil of ignorance (e.g. Brennan and Buchanan, 1985, pp. 28–31).

individuals can be presumed to be mutually beneficial. Thus, ordinary market transactions are a paradigm case of joint actions that benefit some individuals without harming others.[6]

A classic analysis by Buchanan (1968), modelled on that of Knut Wicksell (1896/ 1958), shows how the principle of voluntary exchange can be extended to public goods. The essential idea is that public goods differ from private ones only in respect of the number of individuals involved in the relevant transactions. A public good can be supplied through a mutually beneficial transaction if the costs of supplying it are allocated among the beneficiaries in such a way that, for each individual, the benefits exceed the costs. Of course, such multilateral transactions are much more difficult to negotiate than bilateral transactions in private goods, and it would unrealistic to expect bargaining between large numbers of individual beneficiaries to be an effective mechanism for supplying public goods. Nevertheless, the idea of voluntary exchange provides a template for contractarian recommendations about the provision of public goods. The aim of such a recommendation is to show how a mutually beneficial transaction can be constructed by combining the supply of a particular public good with an appropriate allocation of the costs between beneficiaries.

Similarly, where specific policy proposals impose harms on particular individuals, contractarian policy recommendations may include compensation payments. The principle of analysing policy proposals in conjunction with compensation payments is standard practice in cost-benefit analysis, in the form of the 'compensation test' or 'potential Pareto improvement criterion'. A proposal satisfies this test if it can be combined with a package of compensation payments such that no individuals are no net losers and some are net beneficiaries. Viewed in the contractarian perspective, a cost-benefit analysis that is structured in this way is a first step in identifying opportunities for mutual benefit.

Some readers may object to what they see as the excessive conservatism of a criterion that requires that losers are always compensated. But it is important to recognise the distinction between the contractarian perspective and the view from nowhere. The contractarian is not claiming that the payment of compensation is a necessary means to achieving the overall good of society, viewed impartially. He is not saying that, in an impartial assessment of the social good, one individual's greater gain never outweighs

[6] This claim has to be interpreted with care. A voluntary exchange of private goods between *A* and *B* does not affect other individuals' *holdings of goods*. Nor does it restrict other individuals' general freedom to trade with willing partners. However, it may affect the terms on which further voluntary trades can be made. For example, if *A* wants to sell a car and *B* and *C* both want to buy one, a trade between *A* and *B* may make it more difficult for *C* to find a willing seller.

another's lesser loss. He is addressing individuals, advising them about how to achieve their separate interests through mutually beneficial agreements. If a policy imposes net losses on some individual, the contractarian cannot tell *her* that it is in *her* interest to accept a loss because others are gaining more. The idea that losers are to be compensated is not a moral assumption of contractarian reasoning; it is another expression of the fundamental idea that that reasoning is addressed to individuals.

Ultimately, the concept of a contractarian recommendation – like that of the benevolent despot, and like that of impartial public reasoning – is only a model. It provides a framework for organising normative ideas about economics. If we economists are to think clearly about our normative recommendations, we need some way of construing politics that allows those recommendations a point of engagement. In other words, we need a model of politics in which there are actors to whom our recommendations can be addressed. Since our recommendations are structured by the logic of economic theory, the model must be one in which the addressees have some reason or motivation to act on recommendations that are structured in this way. And, obviously, if the model is to be useful, it must capture significant features of real politics. I suggest each of the three models – the benevolent despot, impartial public reasoning, and contractarian reasoning – is a viable option.

Each model isolates a different aspect of the complex reality of politics in a way that allows economists' recommendations to gain traction. In real politics, there are decision-makers – presidents, ministers of state, senior public servants – who sometimes have both discretionary power and the desire to use this power for the social good. The model of the benevolent despot provides a stylised representation of this form of *politics as executive action* and of the corresponding role of normative economics. In real politics, too, there are arenas of debate about the public good where the participants – parliamentarians, academics, religious thinkers, journalists – strive to deploy impartial and reasoned argument. The model of public reasoning provides a stylised representation of this form of *politics as debate*, allowing a different point of engagement for economists' recommendations. The contractarian model represents politics in a yet another manifestation – *politics as negotiation*. In real politics, there are parties and interest groups whose preferences are neither fully aligned nor completely opposed; politics provides a space in which acceptable compromises are negotiated and mutually beneficial policy packages are identified. The contractarian model allows normative economic reasoning to be brought to bear on this kind of politics.

To some extent, the choice between these models comes down to horses for courses: which model is most useful depends on the problems with which one is dealing. But I think that there is more to the choice than this. Most readers will probably agree that democratic politics, as actually practised, involves elements of executive action *and* of debate *and* of negotiation. They will probably also agree that each of these elements has some legitimate place in democratic politics. But the relative importance of these elements – the importance that they do have, and they importance that they ought to have – is a matter of political judgement and opinion. I would not be writing this paper if I did not believe negotiation to be a major part of what politics is, and of what it should be.

4. A mere modus vivendi?

Political philosophers sometimes suggest that negotiation is an inferior substitute for public reasoning. The thought is this: Sometimes there may be no practical alternative to finding compromises between conflicting interests, but the ideal to which political actors should aspire is a consensus that can be justified by impartial reasoning from premises that they all endorse. With reference to the basic political structure of a democratic society, Rawls (1993) distinguishes between a *modus viviendi* and a *consensus*. A political structure is a modus vivendi when it is viewed by each individual simply as an agreement which, all things considered, works to his or her benefit; among the things that are being considered is the necessity that others can see the agreement as working to *their* benefit. Beyond this recognition of mutual advantage, there is no deeper moral commitment. In contrast, the political structure is supported by a consensus if it rests on moral principles to which everyone is committed. Rawls's concept of a modus vivendi is contractarian in the sense in which I use the term: it fits with a model of politics as negotiation. Similarly, his concept of consensus fits with a model of politics as public reasoning. In Rawls's account of political liberalism, different individuals may subscribe to different (but 'reasonable') comprehensive religious, philosophical, and moral doctrines, but the principles that underlie the political structure are at the intersection of these doctrines – that is, they are the subject of an *overlapping* consensus. All individuals can affirm those principles 'from within their own comprehensive view'; in doing so, each individual 'draw[s] on the religious, philosophical and moral grounds' provided by her own comprehensive view (pp. 15, 147). Rawls argues that a democratic society would have greater stability if its political structure were supported by an overlapping consensus rather than being viewed merely as a modus vivendi.

As a hypothetical proposition, Rawls's claim is perhaps unexceptionable. But negotiating a modus vivendi might still be a more effective way of securing a stable society than trying to create a moral consensus through public reasoning. Suppose you share Rawls's commitment to the principles of political liberalism. Suppose your own comprehensive moral views provide you with what you believe to be sound reasons for affirming those principles. Further, suppose those moral views satisfy Rawls's criterion of reasonableness. Then, naturally, you would *wish* it to be the case that your fellow-citizens subscribed to comprehensive moral views that provided them with what they believed to be sound reasons for affirming the same political principles that you affirm. But wishing something does not make it true. What if other people's comprehensive moral views *don't* give moral support to political liberalism? You may think you have found a philosophical argument that shows that all *reasonable* moral views support political liberalism, but what if other people's beliefs are (as it seems to you) *un*reasonable? Or what if other people don't find your philosophical argument convincing?

When engaging in philosophical argument, it is all too easy to slip from proposing some principle as the *potential* subject of a moral consensus to assuming that that principle *is* supported by such a consensus. A telling example of this slippage can be found in Martha Nussbaum's (2000) development of Amartya Sen's 'capabilities approach' (which in turn is a central component of Sen's account of justice, discussed in Chapter 2). Nussbaum draws up a list of 'central human capabilities' which, she maintains, should be guaranteed to every individual. She proposes that this guarantee should be one of the fundamental principles of political liberalism that, in Rawls's sense, are supported by an overlapping consensus. However, she recognises that in some traditions of religious thought, the denial of some of these capabilities (particularly their denial to women) is treated as a moral requirement. One might have expected her to conclude that, however much she might wish otherwise, public reasoning has not led to a consensus in favour of her list. Instead, her response to religious objections is:

> Given that the religion has agreed to sign on to a constitution of a certain type, it will have to figure out how to square this 'overlapping consensus' on public political matters of basic justice with the rest of what it teaches. (p. 232)

But of course the religions that object to Nussbaum's argument have *not* signed on to a constitution that guarantees her list of capabilities. If she really wants the moral principles

that she affirms to be supported by a consensus, she has to convince everyone else – not tell them that it is their job to work out why she is right.

The lesson I draw from this example is that if you are concerned about the stability of social institutions that you value, you need to find arguments in support of those institutions that your fellow-citizens *in fact* find persuasive. There are no short-cuts. If your fellow-citizens are not persuaded, it is no help to say that they ought to be. Once you recognise that the objective is to find arguments that others will accept, the advantage of deploying arguments that engage with their own interests, as they themselves perceive them, become obvious. In the world as it really is – a world in which people do not easily agree with one another on political, economic or moral questions, and in which failures to agree can all too quickly escalate into conflicts from which everyone loses – there are many worse things than a modus vivendi.

5. Why a contractarian cannot be a paternalist

The distinction between the contractarian perspective and the view from nowhere is particularly significant in relation to questions about paternalism. Suppose that, in some domain of economic life, individuals appear to be making choices that are not in their own best interests, perhaps because of deficient information, faulty reasoning, lack of attention or failures of self-control. Suppose too that these choices are neither beneficial nor harmful to others. How should such choices be viewed in normative economics? Is it the job of economists to propose ways of aligning individuals' private choices more closely with their interests, and if so, what kinds of proposals should be considered and to whom should they be addressed?

It should go without saying that any proposals should be made with decent humility. We should not assume (as economists too often do) that the axioms of conventional decision and game theory are uncontestable standards of human rationality. We should recognise that contraventions of those axioms might reveal limitations of the theory's conception of rationality rather than inadequacies on the part of the people whose choices the theory is failing to explain. For example, the independence and transitivity axioms of expected utility theory were once regarded as self-evidently valid principles of rationality, which would be violated in real-world choices only through error. But when experimental investigations of decision-making behaviour found robust and systematic deviations from the predictions of

expected utility theory, decision theorists began to realise that the traditional axioms reflected implicit assumptions about human psychology that might in fact be false. New theories of choice under uncertainty were then developed which took account of previously-overlooked emotions such as regret, disappointment, ambiguity aversion and loss aversion, showing how violations of the traditional axioms might be rational responses to such emotions.

We should also recognise that the traditional theory of rational choice is concerned only with what Herbert Simon (1978) calls *substantive rationality* – the actual fit between a person's decisions and her objectives. It does not explain the process of reasoning by which this fit can be achieved. A process of reasoning has *procedural rationality* to the extent that it can achieve a satisfactory degree of fit without exceeding the cognitive capacities of the reasoner. Systematic violations of substantive rationality may sometimes be unavoidable consequences of decision-making heuristics that score highly on the criterion of procedural rationality.

But suppose that all due humility has been shown. Suppose that I, as a behavioural economist, am dealing with a case in which, in my judgement, individuals are not acting in their own best interests. As far as I can see, those individuals are not pursuing genuine interests that my theoretical framework has failed to represent. Nor are they acting on heuristics which, all things considered, are well-adapted to the decision problems they face. They are simply making mistakes. What then?

On the face of it, the obvious answer is that, if I feel some concern about these mistakes, I should address my concerns *to the individuals themselves*. Take an analogy from epidemiology – a science which, like economics, deals with issues of individual behaviour and public policy. Consider an epidemiologist who discovers a statistically significant causal relationship between consumption of some common food product and the prevalence of some illness. An obvious next step is for her to make her findings public in such a way that (perhaps through the mediation of other health professionals) potential consumers of the product are informed. As the case of smoking illustrates, the dissemination of information about health risks can precipitate major shifts in consumption patterns – shifts that may begin well before significant public policy interventions are seen as politically feasible. Indeed, some degree of risk awareness on the part of private individuals may be a precondition for successful public intervention. So there is nothing obviously absurd in

thinking that the role of a professional economist might include telling the general public how to avoid decision-making errors.

Given that economists often characterise their discipline as the science of rational choice, one might expect them to recognise the potential value of helping individuals to make better decisions in their private lives. An example of how normative economics can be oriented in this way can be found in the work of one of the pioneers of neoclassical economics. Philip Wicksteed's *Common Sense of Political Economy* (1910/1933) was one of the first attempts to express the theoretical innovations of neoclassical economics in plain prose rather than the mathematics of calculus. Wicksteed presents economics as a study of the 'general laws of the administration of resources' – that is, the principles of optimisation – and insists that these laws apply 'from end to end of life' (p. 159). Although much of his analysis depends on the assumption that individuals act on consistent preferences, he acknowledges that the art of rational decision-making 'by no means looks after itself' (p. 93). In a chapter entitled 'Economical administration and its difficulties', he gives the reader practical advice on how to avoid common mistakes in decision-making. These mistakes include a surprising number of phenomena that have since been investigated by behavioural economists, including the sunk cost effect, failures of self-control, part-whole inconsistencies and bad-deal aversion.[7]

Wicksteed's concern with promoting rationality in private life can still be found in the teaching of economics, where there is an informal tradition of asserting, to the satisfaction of both teacher and student, that people who understand economics are capable of making better decisions than those who don't. (I remember, as an undergraduate student of economics in the late 1960s, learning that bygones are bygones and feeling superior to those non-economists who succumbed to the sunk cost fallacy.) But the application of rational choice theory to private decision-making has not been taken very seriously as a branch of normative economics. In its respectable forms, normative economics has almost always been addressed to *public* decision-makers.

This orientation is perhaps understandable when it is taken by economists who model individuals as ideally rational agents. Such economists are used to thinking about individuals – admittedly, imaginary ones – who have no need for advice about how to make

[7] Wicksteed discusses these effects on pp. 93, 118, 122 and 33 respectively. The first three of these effects will be familiar to most readers of behavioural economics. Bad-deal aversion is a form of reference-dependence in preferences, considered by Thaler (1985) and analysed more formally by Isoni (forthcoming).

better decisions. But it is surely odd that this approach has been carried over to behavioural economics. Of course, an economist who works as a paid consultant has to address herself to whoever pays her, and academic economists are perhaps more likely to be consulted by government agencies than by private individuals. But I am thinking about the huge body of work in normative economics that is not written to meet the demand of any particular client, but is presented at academic conferences and published in academic books and journals. Although the authors of such work may refer to the 'policy implications' of their research, there is usually no actual policy-maker waiting to put these implications into effect. No doubt the authors hope that their findings will eventually filter into the consciousness of some politician or public official, but the idea that normative economics is addressed to a public decision-maker is, as I said in Section 1, no more than a literary convention. This convention seems rather out of place when what are being discussed are (supposed) mistakes in decisions that are made by private individuals and that do not affect anyone else. At any rate, something is clearly wrong if economists think that their response to the discovery of mistakes in individual decision-making *must* take the form of a recommendation about public policy. Advising individuals on how pursue their own interests in their private lives is a natural counterpart to advising them about how to pursue common interests through agreement. In other words, it is a natural counterpart to the contractarian approach.

But what if we are dealing with a mistake which, although made by a private individual, is partly attributable to some feature of that individual's environment that is under the control of some commercial firm or public agency? This is a central issue in the literature of soft paternalism. Richard Thaler and Cass Sunstein (2008) use the term *choice architecture* for the infrastructure associated with decision problems, and suggest that the professional role of behavioural economists should include acting as, or as advisers to, choice architects – that is, the designers of this infrastructure. Thaler and Sunstein argue that one feature of well-designed choice architecture is that it steers or *nudges* the chooser towards the choices that are in her best interests.

One of their examples of this kind of nudging is the design of cash machines. To withdraw cash from a machine, the customer must first insert a bank card. There is a risk that, through lack of attention, she will forget to retrieve her card. The tendency to make this mistake is augmented by the psychological salience of the money relative to the card: it is easy to think that one's interaction with the machine is closed by taking the money. If an economist or psychologist becomes aware that this is a significant problem, it would

certainly be a sensible response to try to alert the users of cash machines to the risk. But another sensible response would be to consider alternative designs of cash machines. It is less likely that anything (money or card) will be left in the machine if the card is returned before the cash is delivered, particularly if the removal of the card is a precondition for the delivery of the cash.

Imagine a time when cash machines delivered the cash before returning the card. Suppose that, at some significant cost, machines can be retrofitted so that this order of operations is reversed. And suppose that, as a behavioural economist, I conclude that this cost is clearly outweighed by the benefit of reducing the frequency of lost cards. If I take the contractarian approach, I can identify a mutually beneficial transaction between customers and banks (or, more accurately, the shareholders who are the banks' owners). My recommendation to each bank is: Retrofit your machines; tell your customers that you have done this; increase the charges to the customers sufficiently to recover the extra costs. My recommendation to each customer is: Patronise banks which use retrofitted machines, even if their charges are slightly higher than those of other banks. Notice that, as is characteristic of contractarian recommendations in general, it is addressed *to individuals together*. Each individual will benefit by acting on the recommendation I make to him, provided that other individuals act on the recommendations I make to them. There is no paternalism in these recommendations. I am advising each customer to recognise her own propensity to error, and hence her interest in paying a premium for good choice architecture. And I am advising the owners of banks that, if customers are willing to pay such a premium, it is in their interest to cater to that demand.

In the case of the cash machine, the relevant choice architecture is supplied by a profit-making firm. What if instead it is supplied by a public agency, financed from general taxation? If, as a contractarian economist, I am to identify a mutually beneficial transaction in this case, it must be between taxpayers. I can advise each individual about whether the benefits she can expect to receive from the redesigned choice architecture exceed the extra costs she will incur as a taxpayer. Again, there is no paternalism: I am advising each individual about her own propensity to error and about what it is in her interest to do about this.

So a contractarian can recommend an individual to make use of types of choice architecture that nudge her away from mistakes that she knows she is liable to make and that she wishes to avoid. He can make this recommendation in relation to a propensity for error

that she was not *previously* aware of. That is, he can say: This is a mistake that you are liable to make; if you want to avoid making it, I recommend this piece of choice architecture. The contractarian might even recommend the individual to make use of a choice architect whom she trusts, just as someone who is building an extension to his house might make use of a real architect. For example, think of how a firm which sells technologically complex products can gain a reputation for good design; a customer might choose to patronise such a firm in the expectation that its products will be easy to use, even though she cannot specify the problems that good designs can overcome.

But what a contractarian economist cannot do is to propose nudging an individual *who does not choose to be nudged*. Such proposals are out of bounds to the contractarian, however much the nudge might seem to be in the individual's interest, and however convinced the economist might be that the individual is making a mistake in not recognising the value of the nudge. The contractarian cannot appeal over the head of the individual to a supposedly more rational self, claiming that the individual *would have chosen* to be nudged, if only she had been better informed, less impulsive, or better able to understand sound reasoning. All of these putative justifications for nudges are paternalistic. They are the kinds of reason that a parent might use to justify her management of a child's behaviour. The parent who tells the child to eat up the vegetables on his dinner plate or to come home before it gets dark will typically say that she is not imposing her own preferences on the child: the behaviour she is demanding is in the child's best interests, and the child would recognise this fact if he were as well-informed and rational as the parent. The paternalism is embedded in the presumption that the parent is entitled to act as the agent of the child's supposed rational self and as the judge of what that self would have chosen.

Why, in the contractarian perspective, is paternalism out of bounds? The answer is *not* that, all things considered, paternalism has undesirable consequences. Nor is it that paternalism violates individuals' rights or compromises their autonomy, and that rights or autonomy have moral value, as viewed from nowhere. It is that, within the contractarian framework, a paternalistic recommendation *lacks a valid addressee*.

Recall that what is at issue is a proposal to nudge some individual, let us say Bill, who is not choosing to be nudged. Recall too that the supposed mistake that Bill is to be nudged away from does not harm or benefit anyone else. To whom can that proposal be addressed? Clearly, not to Bill himself: if he were being addressed, the recommendation would be that he should *choose* to be nudged. It must be addressed to someone else, whose

relationship to Bill is that of guardian to ward. But contractarian recommendations are not addressed to imagined benevolent despots or to self-appointed guardians. They are addressed to individuals as the directors of their own lives, advising those individuals about how to pursue their own interests. Paternalistic proposals are not recommendations of this kind; in a contractarian analysis they are simply out of place. One might say that they are *ultra vires*, not properly on the agenda for contractarian discussion.[8]

In the contractarian perspective, the question of whether or not the supposed beneficiary of a nudge – the *nudgee* – has chosen to be nudged is fundamental. But if one takes the view from nowhere, this question is much less significant. The impartially benevolent spectator who takes this view is concerned with the good of each individual, all things considered. So when she thinks about a proposal to nudge Bill, she asks herself whether that nudge would be good for Bill; and that judgement is ultimately hers, not Bill's. There is nothing improper in her judging that it would be good for Bill, even though Bill thinks otherwise. And such a judgement has an addressee: it can be addressed to an imagined benevolent despot, as welfare judgements are in traditional welfare economics, or it can be a contribution to a process of public reasoning about the social good.

As I pointed out in Section 1, the literature of soft paternalism takes the view from nowhere. So it is perhaps not surprising that, in this literature, questions about whether individuals choose to be nudged are not given much attention, or receive only casual answers. Thaler and Sunstein's (2008) advocacy of libertarian paternalism illustrates this point.

Thaler and Sunstein start from the proposition that 'individuals make pretty bad decisions – decisions they would not have made if they had paid full attention and possessed complete information, unlimited cognitive abilities, and complete self-control'. Nudges are designed to counteract these imperfections of individual decision-making. Thaler and Sunstein concede that, in proposing nudges, they are being paternalistic: 'The paternalistic aspect [of libertarian paternalism] lies in the claim that it is legitimate for choice architects to try to influence people's behavior in order to make their lives longer, healthier and better'.

[8] This bald claim needs some qualification in respect of children and the mentally incompetent (such as people with advanced Alzheimer's disease). If we are to use a normative framework based on voluntary contract, we must recognise that at least some of the interests of children and the mentally incompetent have to be looked after by agents who act in the role of guardian or trustee. In these cases, contractarian recommendations *can* properly be addressed to guardians. How to draw the line between the domains of responsible choice and guardianship is an important problem for normative economics, and particularly so for its contractarian form.

The libertarian aspect of libertarian paternalism is the principle that choice architects must not significantly obstruct individuals' freedom of choice – they must rely on nudges. I shall call this the *free choice condition*. The idea is to take advantage of what behavioural economics has shown to be the malleability of people's preferences. Well-designed choice architecture nudges people towards the choices that are in their best interests, while leaving them free to choose otherwise if they really want to. Notice that the free choice condition sets limits to the kinds of paternalistic policies that can be recommended, but it is compatible with paternalism within those limits (which is why Thaler and Sunstein can deny that the term 'libertarian paternalism' is an oxymoron).

Thaler and Sunstein insist that their recommendations are designed to 'make choosers better off, *as judged by themselves*' (p. 5, italics in original). I take it that the italicised clause, which is repeated with minor variations at other places in their book (e.g. pp. 10, 12, 80), is intended to signal that Thaler and Sunstein's nudges will be designed to steer each individual towards the decisions that she would have made, had she been *perfectly rational* – that is, had she paid full attention and possessed complete information, unlimited cognitive abilities and complete self-control. This clause may seem to make Thaler and Sunstein's approach more benign than traditional forms of paternalism, but appearances here are deceptive.

Determining what a person would choose, were she perfectly rational, is not just a matter of discovering given facts about her. The concepts of full attention, perfect information, unlimited cognitive ability and complete self-control do not have objective definitions; they are inescapably normative. Just about any intervention that a paternalist sincerely judges to be in the individual's best interests can be justified in this way if the paternalist is allowed to define what counts as attention, information, cognitive ability and self-control. The claim that the paternalist is merely implementing what the individual would have chosen for herself under ideal conditions is a common theme in paternalistic arguments, but should always be viewed with scepticism.

Even if Thaler and Sunstein's concept of perfect rationality could be defined objectively, there might still be no determinate answer to the question of what an individual would have chosen, had he been perfectly rational. Thaler and Sunstein seem to be assuming that inside every imperfect human being there is a neoclassical rational agent – that, deep down, each of us has coherent preferences, of the kind that economic theory has traditionally postulated, and that these can be found by stripping away specific failures of

rationality. But the experimental evidence on which behavioural economics is grounded does not support this assumption. I conclude that the 'as judged by themselves' clause is more of a rhetorical flourish than a genuine restriction on paternalism.

When justifying specific proposals for nudging, Thaler and Sunstein sometimes claim more than that nudgees will be made better off, as judged by themselves (or rather, as they would judge, were they perfectly rational). Thaler and Sunstein make the further claim that the nudgees *want* to be nudged. If this claim were true, nudging would not be paternalistic, and might be justified on contractarian grounds. But typically the claim is made in vague terms and with little supporting evidence. Thaler and Sunstein sometimes appeal to the 'New Year's resolution test'. For example, in support of nudging individuals towards healthier lifestyles: '[H]ow many people vow to smoke more cigarettes, drink more martinis, or have more chocolate donuts in the morning next year?' (p. 73). More substantially, in support of nudging individuals to save more, they cite survey evidence that two-thirds of employees describe their savings rate as 'too low' while only one per cent describe it as 'too high'. Such statements are, they say, 'not meaningless or random' (p. 107). That is true, but the test that has been satisfied is not exactly stringent. One might have hoped for a criterion that could discriminate between the New Year's resolutions that many of us make without seriously expecting (or even trying) to keep and genuine personal commitments that fail only under intense psychological pressure.

The idea that nudgees want to be nudged in just the directions that Sunstein and Thaler propose to nudge them is supported by an implicit assumption about expertise. The assumption is not merely that nudgees are willing to defer to the expertise of choice architects; it is that Sunstein and Thaler's own scientific judgements constitute expertise, *as judged by nudgees*. In relation to many of the nudges that Sunstein and Thaler propose, that assumption seems implausible. Take the case of diet. Think of all those people who consciously try to manage their diets in the interests of health or good looks (but without forgetting how many other people never give this a second thought, and have no desire to change their behaviour). A typical dieter will be acting on some amalgam of the vast amount of dietary advice that is disseminated in television programmes, newspaper reports, magazine articles, popular books and advertisements. As viewed by professional epidemiologists, some of this advice is clearly grounded in good science, some is scientifically controversial, some is harmless crackpottery, and some is downright dangerous. But to each dieter, the advice on which he acts *is* expertise. Epidemiologists

may agree that some popular dietary guru is no more than a quack, but to the guru's followers she is a scientific authority. An epidemiologist might reasonably claim that dieters would benefit from help in choosing their advisors, if that help were based on the expertise of epidemiologists like themselves; but the question at issue is whether *the dieters themselves* believe that they are in need of such help. The fact that quackery can coexist with widely disseminated official health advice suggests that in many cases the answer is 'No'.

Reading between the lines of Sunstein and Thaler's text, I sometimes detect a suggestion that precision in defining the 'as judged by themselves' condition isn't really required, since individuals are *only* being nudged. For example, after appealing to the New Year's resolution test and after conceding its obvious limitations, Thaler and Sunstein say that they interpret statements of the form 'I should be saving (or dieting, or exercising) more' as implying that the individuals who make them 'are open to a nudge' (a usefully vague notion) and 'might even be grateful for one' (p. 107). In other words, they do not claim that such self-critical statements provide evidence that the individuals who make them *do* want to be nudged, but only they *might* want to be nudged; and that, it seems, is good enough. The underlying thought is that if the free choice condition is satisfied, there cannot be any serious objection to paternalism.

This thought is made explicit in an earlier paper, in which Sunstein and Thaler (2003a) consider the objection that autonomy has moral value, and that 'people are entitled to make their own choices even if they err'. Their response is:

> We do not disagree with the view that autonomy has claims of its own, but we believe that it would be fanatical, in the settings we discuss, to treat autonomy, in the form of freedom of choice, as a kind of trump, not to be overridden on consequentialist grounds. ... [W]e think that respect for autonomy is adequately accommodated by the libertarian aspect of libertarian paternalism. (p. 1167, note 19)

Notice that the objection to which Sunstein and Thaler are responding is another view from nowhere. They are imagining a critic who maintains that autonomy is a component of individual well-being, and so ought to be included in any assessment of what is good, all things considered. They 'do not disagree' with this general idea, but think that only a fanatical libertarian would appeal to it as an objection to the sort of nudges they are proposing. When an individual's own choices – say, through excessive drinking or over-

eating – are so much in error that they seriously impair his health, how can the effects on his autonomy of a mere nudge outweigh the prospective benefits in the form of better health?

If one takes the view from nowhere, this argument has some force. But it is not an argument against the contractarian position. The contractarian does not claim that unchosen nudges (that is, nudges that are not chosen by the nudgee) are bad, all things considered, but only that they cannot be recommended to the nudgee. From long experience of giving talks on this topic, I know that many economists and philosophers *do* think that the contractarian position is fanatical. A typical questioner will describe some case in which a mild but unchosen nudge would be very beneficial to the nudgee (as judged by the questioner). Perhaps the nudgees are morbidly obese, and the nudge is a government policy that will make unhealthy fast food less readily available. The questioner asks me: What would you do in this case? To which my reply is: What do you mean, what would *I* do? What is the imaginary scenario in which I am supposed to be capable of doing something about the diets of my morbidly obese fellow-citizens?

If the scenario is one in which Robert Sugden is in a roadside restaurant and a morbidly obese stranger is sitting at another table ordering a huge all-day breakfast as a mid-afternoon snack, the answer is that I would do nothing. I would think it was not my business as a diner in a restaurant to make gratuitous interventions into other diners' decisions about what to eat. But of course, this isn't the kind of scenario the questioner has in mind. What is really being asked is what I would do, *were I a benevolent despot*. My answer (which, I must confess, does not usually satisfy the questioner) is that I am not a benevolent despot, nor the adviser to one. As a normative economist, I am not imagining myself in either of those roles. I am advising individuals about how to pursue their common interests, and there is no common interest in unchosen nudges.

6. The Four Alls

Some readers may by now have been persuaded of the internal coherence of the contractarian approach, but still be surprised that anyone would *want* to think about normative economic issues in this way. It may be true (they may think) that paternalism is out of bounds to contractarian analysis, but isn't that just a symptom of the deficiencies of that approach? I will end this chapter with an attempt to express how, if one looks at the world in a certain way, the contractarian approach can be seen as attractive.

Thaler and Sunstein devote a chapter of *Nudge* to the issue of retirement savings. The content of this chapter is summarised in the final paragraph:

> Saving for retirement is something that Humans [as contrasted with ideally rational agents] find difficult. They have to solve a complicated mathematical problem to know how much to save, and then they have to exert a lot of willpower for a long time to execute this plan. This is an ideal domain for nudging. In an environment in which people have to make only one decision per lifetime, we should surely try harder to help them get it right. (2008, p. 117)

Look at the final sentence. Thaler and Sunstein are telling their readers that *we* should try harder to help *them* get their decisions right. But who are the 'we' and who are the 'they' here? What 'we' are supposed to be doing is designing and implementing choice architecture which nudges individuals to save more for retirement; so presumably 'we' refers to government ministers, legislators, regulators, human resource directors and their respective assistants and advisers; 'they' are the individuals who should be saving. As an expert adviser on the design of occupational pension schemes, Thaler is certainly entitled to categorise himself as one of the 'we'. But where do his readers belong? Very few of them will be in any position to design savings schemes, but just about all of them will face, or will have faced, the problem of saving for retirement. From a reader's point of view, Thaler and Sunstein's conclusion would be much more naturally expressed as: *They* should try harder to help *us* get it right. Thaler and Sunstein are writing from the perspective of *insiders* to the public decision-making process: they are writing as if they were political or economic decision-makers with discretionary power, or their trusted advisors. And they are inviting their readers to imagine that they are insiders too – that they are the people in control of the nudging, not the people who are being nudged.

I suggest that the benevolent despot model appeals to people who like to imagine themselves as insiders in this sense. By imagining yourself into a suitable insider role, you can forget about all the real obstacles that lie between your having (what you believe to be) a good idea about how other people's welfare might be improved and there being a public decision to implement that idea. If you have been trained as an economist, you can imagine advising a decision-maker who shares your belief in the importance of economics and who has the good sense to consult sound economists such as yourself. You do not have to ask whether real decision-makers would want to take your advice. Nor do you have to ask whether other people's ideas about how *your* welfare might be improved – ideas that *they* believe to be good, but you perhaps don't – might get implemented instead.

The public reasoning model has something of the same appeal to people who like to imagine themselves as insiders in a different sense. That model invites you to imagine a public discussion in which each participant presents reasoned arguments in support of his or her judgments about the overall good, and all try to reach agreement about the validity and force of these arguments. If you have been trained as a philosopher, you have professional expertise in reasoned argument. You can imagine having a prominent role in a public discussion, presenting arguments which carry the day by virtue of their philosophical merits. It is easy to forget that other people's arguments, based on reasons which they believe to be sound but you don't, might prove more persuasive.

In contrast, the contractarian approach appeals to people who take an outsider's view of politics, thinking of public decision-makers as agents and themselves as principals. The sort of person I have in mind does not think that he has been unjustly *excluded* from public decision-making or debate; he is more likely to say that he has (what for him are) more important things to do with his time. He does not claim to have special skills in economics or politics, and is willing to leave the day-to-day details of public decision-making to those who do – just as he is willing to leave the day-to-day maintenance of his central heating system to a trained technician. But when public decision-makers are dealing with his affairs, he expects them to act in his interests, as he perceives them. He does not expect them to set themselves up as his guardians.

This way of thinking about politics is encapsulated in a traditional British inn sign, the sign of the Four Alls. The sign is divided into quarters, on the model of an heraldic shield. The first quarter shows a picture of a King, with the words 'I rule all'. The second shows a soldier: 'I fight for all'. The third shows a parson: 'I pray for all'. The fourth shows a farmer, with the words 'I pay for all'. The sign expresses the farmer's view of public affairs. The farmer, I take it, recognises the value he derives from the activities of the government, the army and the church. He does not pretend to possess the particular skills that those activities require, and has no particular wish to do so: he has his own skills, which are at least as valuable in the overall scheme of things. But he expects the King, the soldier and the parson to remember that it is his taxes that pays for their work. He does not defer to them: he is their employer.

References

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

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

Brennan, Geoffrey and James M. Buchanan (1985). *The Reason of Rules*. Cambridge: Cambridge University Press.

Buchanan, James M. (1954). Individual choice in voting and the market. *Journal of Political Economy* 62: 334–43.

Buchanan, James M. (1968). *The Demand and Supply of Public Goods*. Chicago: Rand McNally.

Buchanan, James M. (1975). *The Limits of Liberty*. Chicago: University of Chicago Press.

Buchanan, James M. (1986). *Liberty, Market and State*. Brighton: Wheatsheaf.

Buchanan, James M. and Gordon Tullock (1962). *The Calculus of Consent*. Ann Arbor: University of Michigan Press.

Camerer, Colin F., 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.

Gauthier, David (1986). *Morals by Agreement*. Oxford: Oxford University Press.

Habermas, Jürgen (1995). Reconciliation through the public use of reason: remarks on John Rawls's political liberalism. *Journal of Philosophy* 92: 109–131.

Harsanyi, John C. (1955). Cardinal welfare, individualistic ethics and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.

Hobbes, Thomas (1651/ 1962). *Leviathan*. London: Macmillan.

Hume, David (1739-40/ 1978). *A Treatise of Human Nature*. Oxford: Oxford University Press.

Isoni, Andrea (forthcoming). Price sensitivity and 'bad-deal' aversion: a new explanation for the WTA/WTP disparity in Vickrey auctions. Forthcoming in *Theory and Decision*.

Loewenstein, George and Peter A. Ubel (2008). Hedonic adaptation and the role of decision and experience utility in public policy. *Journal of Public Economics* 92: 1795–1810.

Mackie, John L. (1977). *Ethics: Inventing Right and Wrong*. London: Penguin.

McQuillin, Ben and Robert Sugden (2011). How the market responds to dynamically inconsistent preferences. Forthcoming in *Social Choice and Welfare*.

Nagel, Thomas (1986). *The View From Nowhere*. Oxford: Oxford University Press.

Nussbaum, Martha (2000). *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press.

Putnam, Hillary (2004). *Ethics without Ontology*. Cambridge, MA: Harvard University Press.

Rawls, John (1971). *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press.

Rawls, John (1993). *Political Liberalism*. New York: Columbia University Press.

Sen, Amartya (1999). *Development as Freedom*. Oxford: Oxford University Press.

Sen, Amartya (2009). *The Idea of Justice*. London: Allen Lane.

Simon, Herbert A. (1978). Rationality as process and as product of thought. *American Economic Review* 68: 1–16.

Smith, Adam (1759/ 1976). *The Theory of Moral Sentiments*. Oxford: Oxford University Press.

Sugden, Robert (2004a). *The economics of rights,cooperation and welfare*. Second edition. Palgrave Macmillan, Basingstoke. First edition 1986.

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

Sugden, Robert (2008). Why incoherent preferences do not justify paternalism. *Constitutional Political Economy* 19 (2008): 226–248.

Sugden, Robert (2010). Opportunity as mutual advantage. *Economics and Philosophy* 26: 47–68.

Sunstein, Cass R. and Richard H. Thaler (2003a). Libertarian paternalism. *American Economic Review, Papers and Proceedings* 93 (2): 175-179.

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

Thaler, Richard (1985). Mental accounting and consumer choice. *Marketing Science* 4: 199-214.

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

Wicksell, Knut (1896/ 1958). A new principle of just taxation. English translation (by James M. Buchanan) of German original. In Richard A. Musgrave and Alan T. Peacock (eds), *Classics in the Theory of Public Finance*, 72–118. London: Macmillan.

Wicksteed, Philip (1910/ 1933). *The Common Sense of Political Economy*. London: Routledge.