CBESS Discussion Paper 12-05

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

by Robert Sugden*

*School of Economics, University of East Anglia

Abstract
This paper compares two alternative answers to the question 'Who is the addressee of welfare economics?' These answers correspond with different understandings of the status of the normative conclusions of welfare economics, and have different implications for how welfare economics should be adapted in the light of the findings of behavioural economics. The conventional welfarist answer is that welfare economics is addressed to a 'social planner' whose objective is to maximise the overall well-being of society; the planner is imagined as a benevolent despot, receptive to the economist's advice. The alternative contractarian answer is that welfare economics is addressed to individuals who are seeking mutually beneficial agreements; a contractarian recommendation has the form 'It is in the interests of each of you separately that all of you together agree to do x'. Each of these answers should be understood as a literary convention which uses a highly-simplified model of politics. I defend the contractarian approach and show that it is less supportive of 'soft paternalism' than is the welfarist approach.

JEL classification codes
D003, D60

Keywords
Welfare economics, behavioural economics, social planner, contractarianism, soft paternalism.

Centre for Behavioural and Experimental Social Science
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/ssf/cbess

The behavioural economist and the social planner:

to whom should behavioural welfare economics be addressed?[1]

Robert Sugden

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

26 June 2012

[1] This paper contains material from 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 that does not depend on strong and implausible assumptions about individual rationality.

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 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; instead, whatever preferences they need to deal with a 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, 2004, 2008, 2010; McQuillin and Sugden, 2012). 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, she somehow rises above her ordinary self and assumes a viewpoint from which she 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, rather odd.

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 how society would look to an *impartially benevolent spectator*.

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? To whom is welfare economics addressed?

The traditional addressee of welfare economics is an entity variously known as 'the 'the policy-maker', 'the government' or 'the social planner'. 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.

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, for Sunstein and Thaler, 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 suggesting that she should commission me to report on the feasibility of a coup, and on whether that would result in an increase in 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*.

I will now present an alternative 'contractarian' model in which there is a different addressee (or as will become clear, addressees). I 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. The contractarian perspective

In the sense in which I will use the term 'contractarian', the most fundamental characteristic of the contractarian 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.

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 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.

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. As a simple 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. 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.

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 the example of *A* and *B* negotiating over the sale of a car. 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; the relevant benchmark is the allocation of resources prior to trade. If the trade takes place in a market with many potential buyers and sellers, it is also reasonable to presume that no

one else's interests are significantly affected. Thus, ordinary market transactions provide a model of joint actions that benefit some individuals without significantly harming others.[2]

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 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

[2] The qualification 'significantly' hides some important issues. A voluntary exchange of private goods between two individuals can have 'pecuniary' externalities on others through its effects on market prices. In a market with many buyers and sellers, these effects are *individually* very small, but it is possible that some individuals incur significant losses as a result of the cumulative effects of other people's voluntary transactions. I discuss the implications of this possibility for contractarian reasoning in Sugden (2012).

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, 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. The model of contractarian reasoning, like that of the benevolent despot, satisfies these conditions.

Each of the two models isolates a particular 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 aspect of *politics as executive action* and of the corresponding role of normative economics. The contractarian model represents politics in a different 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 more 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 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.

3. 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 I, as a behavioural economist, am dealing with a case in which, in my judgement, individuals are not acting in their own best interests, perhaps because of deficient information, faulty reasoning, lack of attention or failures of self-control. Suppose too that these individuals' choices are neither beneficial nor harmful to others. How should I respond?

On the face of it, the obvious answer is that if I feel some concern about these faulty decisions, 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. Some traces of this way of thinking can indeed 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. But 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 better decisions. But it is surely odd that this approach has been carried over to behavioural economics. The literary convention of addressing normative economics to a public decision-maker 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. 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. Thaler and 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 her house might make use

of a real architect. 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 its value. 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*. 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.[3]

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

[3] 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.

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 someone, she asks herself whether that nudge would be good for the nudgee; and that judgement is ultimately hers, not the nudgee's. There is nothing improper in her judging that it would be good for the nudgee, even though the nudgee thinks otherwise. 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'.

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 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 what they recognise as their common interests. Unless individuals themselves wish to license others to act as their guardians, there is no common interest in paternalism.

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.

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.

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.

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.

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

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

Sugden, Robert (2004). 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.

Sugden, Robert (2012). The market as a cooperative endeavour. Forthcoming in *Public Choice*.

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 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.