On the Possibility of an Anti-Paternalist Behavioural Welfare Economics

Johanna Thoma
London School of Economics and Political Science

July 13, 2021

Abstract

Behavioural economics has taught us that human agents don't always display consistent, context-independent and stable preferences in their choice behaviour. Can we nevertheless do welfare economics in a way that lives up to the anti-paternalist ideal most economists subscribe to? I here discuss Sugden's powerful critique of most previous attempts at doing so, which he dubs the 'New Consensus', as appealing to problematic notions of latent preference and inner rational agency. I elaborate on a fundamental rethinking of the normative foundations of anti-paternalist welfare measurement that often remains implicit in the behavioural welfare economics literature Sugden discusses, but which is required to make these accounts minimally plausible. I argue that, if we go along with this rethinking, Bernheim and Rangel's (2007, 2009) choice-theoretic framework withstands Sugden's criticism. Sugden's own, more radical proposal is thus under-motivated by his critique of the 'New Consensus'.

Keywords: Behavioural economics; Welfare economics; Anti-paternalism; Preference Purification; Choice

1 Introduction: Anti-Paternalist Welfare Measurement

Many welfare economists take as a starting point of their analysis the idea that, when evaluating the effects of institutions, policies and interventions on individuals, they should be deferential to those individuals' *subjective interests*, that is, to their own views about what is good for them, or what they want to do or be done to them. Economists should avoid imposing their own, or a third party's idea of what is good for a person. Call this general ambition the ideal of 'anti-paternalist' welfare measurement.[1]

For the purposes of this paper, I understand 'welfare' as a generic metric for taking into account the effects on individuals when conducting an overall evaluation of an institution, policy or intervention. Philosophers of economics sometimes assume that 'welfare' as used by welfare economists must be understood as just another word for what philosophers call 'wellbeing', that is, a measure of what is truly good for a person. They then further take welfare analysis more generally to be engaged in a consequentialist project aiming to improve people's true wellbeing. On that interpretation, commitment to the anti-paternalist ideal must be motivated either by commitment to a subjective account of well-being, that is, the view that an individual's subjective conative attitudes, such as her preferences, desires, plans, or commitments, constitute what is good for that person, or by the view that subjective interests are the best evidence we have of what is truly good for somebody.

However, if we take a broader view of welfare as not necessarily equated with well-being, the ideal of anti-paternalist welfare measurement can also be justified on non-consequentialist and broadly liberal grounds. For instance, we might think that deference to people's subjective interests is required as a way of showing respect for their autonomy, or of exercising liberal neutrality amongst different conceptions of the good life – even while acknowledging that a person might be mistaken about what is in fact good for her.

One important liberal line of justification for the anti-paternalist ideal can be found in the social contract tradition, wherein the rules and institutions that govern our lives must be justifiable to each citizen. Robert Sugden's (2018) book *The Community of Advantage* falls within this tradition, and more specifically the contractarian one. Sugden argues that not only should the state or policy-maker aim to remain neutral between different conceptions of the good life in order to remain justifiable to all, the state or policy-maker is not even the proper addressee of the recommendations of the welfare economist. Rather, each citizen needs to be shown that some institution or proposed policy is in her interests. And she can only be shown this if we appeal to her subjective interests.

[1] I do not mean this to be a full characterization of anti-paternalism in general. For a full definition of paternalism in terms of a lack of deference to an agent's subjective interests, see Haybron and Alexandrova (2013).

I will take the anti-paternalist ideal for granted in the following without committing to any of its potential consequentialist or non-consequentialist justifications. I will instead focus on the question of its implementation in welfare economics. As I described it above, the anti-paternalist ideal does not commit us yet to a specific account of which of an agent's attitudes constitute her subjective interests, and how we formally represent them. The orthodox approach to welfare economics formally represents agents' interests, and thus welfare, with a preference relation – a binary relation over the objects of choice – that is derived from their observed choices, and assumed to be what Sugden calls 'integrated' (p.7): consistent according to the axioms of standard rational choice theory, stable, and context-independent. Given this formal representation, it is tempting to also take the mental attitudes that are often assumed to correspond to such preferences – all-things-considered comparative evaluations of the objects of choice – to be the attitudes that *constitute* the agent's subjective interests. On this view, the orthodox approach is not only *formally* preference-based, it is also *normatively* preference-based: the preferences that feature in economic models are direct representations of the mental attitudes that constitute an agent's subjective interests and that therefore serve as the normative basis of anti-paternalist welfare measurement.

An important challenge for orthodox welfare economics has come from behavioural economics, which provides us with ample evidence that real agents do not always exhibit choice behaviours that allow for representation in terms of an integrated preference relation. They regularly violate the standard axioms of rational choice, and exhibit different choice behaviours in different contexts, where the change in context does not seem to intuitively be choice-relevant. For instance, to use an example much discussed in the context of behavioural welfare economics, the placement of food items in a cafeteria line may systematically affect what items are chosen, even though this placement is not something most people intrinsically care about. How, then, should welfare economics proceed to measure the welfare of agents who exhibit inconsistent or context-dependent preferences? In particular, how can it still implement the anti-paternalist ideal?

Sugden's book, and in particular Chapter 4, which builds on joint work with Gerardo Infante and Guilhem Lecouteux (Infante et al. 2016), does a remarkable job at both characterizing the common core of many recent attempts at answering this challenge, as well as presenting a powerful critique of this common core. At the heart of what Sugden calls the 'New Consensus' in behavioural welfare economics is the idea that, while agents may not exhibit integrated preferences in their behaviour, we can still ascribe to them 'latent' preferences that are integrated. It is these 'latent' preferences that we can then use as a measure of their welfare.

Sugden's critical analysis of the New Consensus serves as motivation for a radically different way of measuring economic welfare, which gives up on the idea of evaluating

objects of choice, such as outcomes or lotteries, in terms of an agent's subjective interests, but rather ranks only option sets. According to what Sugden calls the 'Opportunity Criterion', defended in Chapter 5, welfare is increased whenever agents receive a strictly expanded option set, as this presents an unambiguous increase in opportunity. Notably, this produces an incomplete welfare ranking, and Sugden presents no other method for ranking opportunity sets that are not strict expansions or contractions of each other. Still, as each agent can see that an unambiguous increase in opportunity is in her continued interests, Sugden argues that this criterion, and an opportunity-based collective criterion defended in Chapter 6, can serve as an appropriate basis for his contractarianism.

In the following, I aim to show that Sugden's criticism of the New Consensus is not sufficient to motivate giving up on evaluating objects of choice in terms of an agent's subjective interests. A crucial question in this debate, as we will see, is whether the approaches Sugden classifies as part of the New Consensus hold on to the idea that welfare measurement is not only *formally* preference-based (by featuring a latent preference relation), but also *normatively* preference-based, taking subjective interests to be constituted (and not merely formally represented) by an integrated preference-like mental attitude. If the answer is 'yes', I argue that the New Consensus can be dismissed fairly quickly, as it faces a dilemma that cannot be plausibly resolved. However, much of Sugden's own critique allows for attitudes that do not directly correspond to integrated preferences to constitute subjective interests. In particular, he appears to allow for desires regarding the attributes of the different objects of choice to play such a role. And in that case, at least one approach Sugden dismisses as part of the New Consensus in fact withstands Sugden's critique, namely a choice-theoretic framework developed by Bernheim and Rangel (2007, 2009), under the revised interpretation recently offered by Bernheim (2016). This framework, I argue, is a plausible way to implement the anti-paternalist ideal in the face of behavioural anomalies.

2 The New Consensus

What unites the 'New Consensus', as characterized by Sugden is the idea that agents who do not exhibit integrated preferences in their behaviour can still be ascribed 'latent' preferences that are integrated. We should then use these 'latent' preferences as a measure of their welfare. The most sophisticated work that clearly exemplifies this New Consensus constructs behavioural models that can accommodate inconsistent choice behaviours, while at the same time estimating parameters that can be plugged into normative models of how the agent should rationally choose which do yield integrated preferences. We can then think of the differences between actual behaviour and the choices the normative model recommends as the result of various biases.

To illustrate, where agents violate the axioms of expected utility theory in contexts of uncertainty, cumulative prospect theory is often taken to be the correct behavioural theory. Bleichrodt et al. (2001) use prospect theoretic models to estimate a utility and probability function from an agent's observed choices, and then use these to construct a normative expected utility model, correcting for the effects of loss aversion and probability-weighting in the original prospect theoretic model, which are treated as biases. The resulting integrated preferences are then used as a welfare measure.[2] Bershears et al. (2008), Köszegi and Rabin (2007), Manzini and Mariotti (2012) and Salant and Rubinstein (2008) also pursue such a reconstructive approach to behavioural welfare economics.

Other authors clearly exemplifying Sugden's New Consensus don't offer a formally rigorous way of reconstructing integrated preferences from the inconsistent, context-dependent or unstable behaviour we observe. But they nevertheless share the conviction that (1) inconsistency, context-sensitivity or instability in observed choice is evidence of some form of error, and that (2) we can still identify error-free, integrated preferences that serve as an appropriate welfare standard in such cases. For instance, Thaler and Sunstein's (2008) libertarian paternalism claims to intervene only so as to help agents achieve what is best for them, 'as judged by themselves' (p.5).[3] Sugden also classifies a choice-theoretic approach to behavioural welfare economics proposed by Bernheim and Rangel (2007, 2009) as part of the problematic New Consensus. I will offer a little more detail on this approach here, because I will later argue that, given the right interpretation, it in fact avoids the problems Sugden raises for the New Consensus, and remains a plausible way of implementing the anti-paternalist ideal in the face of behavioural anomalies.

Formally, Bernheim and Rangel's approach starts by identifying 'generalized choice situations' consisting both of a set of objects of choice for an agent, as well as a set of ancillary conditions. Having recorded an agent's choices in various such 'generalized choice situations', the analysis proceeds in two steps: First, out of the choices we have observed, we identify the ones that 'merit deference', and throw out the ones that don't. Here, we should only throw out those choices we can clearly identify as mistakes. For instance, if we have compelling evidence that a vegetarian cafeteria customer both has a sweet tooth and does not know that mince pie is in fact a meat-free sweet pie, we should disregard her choices of inferior sweet snacks over the mince pie.

In the second step, we conduct welfare analysis by identifying and exploiting the coherent aspects of those choices that merit deference. The key welfare criterion for Bernheim and Rangel is the notion of 'unambiguous choice'. Informally, 'we say that one alternative is unambiguously superior to another if and only if the second is never

[2] See my Thoma (2021) for an extended critical discussion of this work.

[3] Also see the earlier Sunstein and Thaler (2003a,b), Camerer et al. (2003) and Le Grand and New (2015) for a similar treatment.

chosen when the first is available.' (Bernheim 2016, p.15) By using unambiguous choice as a welfare criterion, we are assuming unambiguous choice tracks welfare, and the anti-paternalist welfare economist should defer to it. If a cafeteria customer never chooses liquorish, no matter where it is placed, then the anti-paternalist should defer to this unambiguous choice of any other snack over liquorish. Where a choice between two options is reversed under different ancillary conditions each of which merit deference, the agent did not unambiguously choose either – for instance, if the customer sometimes picks chocolate over bananas and sometimes bananas over chocolate depending on where the items are placed. According to Bernheim and Rangel, this ambiguity should not be resolved. Rather, we should simply accept that it is indeterminate which of the options is more in the agent's subjective interest to receive. Where all of an agent's choices are consistent, this approach generalizes to standard revealed preference welfare analysis.

While Sugden acknowledges the greater permissiveness of this approach, he nevertheless classifies it as alike enough to the New Consensus to be dismissed on the same basis: 'Although Bernheim and Rangel do not assume that context-independent latent preferences always exist, their approach yields welfare rankings only for those pairs of objects for which revealed preferences, after purification, are context-independent.' (p.58) The thought is that this approach features a welfare measure that is context-independent, stable, and consistent, except, and unlike the approaches presented before, for the fact that it is incomplete. Sugden appears to interpret unambiguous choice as playing the same role as latent preferences in the other approaches of the New Consensus, except for its incompleteness.

3 A Dilemma for Preference-Based Welfare Measurement

Is what Sugden calls the 'New Consensus' tenable from the perspective of the anti-paternalist ideal for welfare economics? Sugden's argument against the New Consensus focusses on the notion of 'latent preference', which he argues to be not fit for purpose. Roughly, Sugden argues, on the one hand, that the New Consensus presupposes a psychologically implausible view of human agents as having an 'inner rational agent' (p.64) who actually has or at least has the capacity to display integrated preferences, and a 'psychological shell' (p.65) that distorts them. Moreover, on the other hand, he claims that the New Consensus relies on the idea that an agent's perfectly rational, strong-willed and well-informed counterpart, an agent Sugden calls 'SuperReasoner' (p.56) would display integrated preferences, and that this, too, is implausible.

How damaging this criticism is, I think, depends on what we take the relation to be between the reconstructed 'latent' preference relation and subjective interest. Like

the orthodox approach, the New Consensus in behavioural welfare economics is formally preference-based: Welfare is formally represented with an integrated preference relation. Suppose we also take it to be *normatively* preference-based, that is, based on the assumption that subjective interest is constituted by mental states that can be directly represented by some integrated preference relation, so mental states that provide an all-things-considered evaluation of the objects of choice. There is then an important ambiguity in the notion of latent preference that Sugden acknowledges at one point (p.79) but does not make central.

The ambiguity is whether latent preferences are meant to represent actual mental states of an agent which don't find expression in her choices for whatever reason – e.g. the actual preferences of her inner rational agent – or whether they are purely hypothetical, representing mental attitudes she should and would have under more ideal circumstances – e.g. the preferences of her perfectly rational counterpart. Several of the authors Sugden discusses in fact equivocate between the two notions. The distinction between these two senses of 'latent preference' is important because it helps us see an apparent dilemma for the New Consensus if it aims to be normatively preference-based: If latent preferences are supposed to be actual mental states, the claim that agents who display behavioural anomalies have latent preferences that are integrated is clearly psychologically unrealistic. But if latent preferences are merely hypothetical, they risk not being appropriately subjective so as to satisfy the anti-paternalist ideal – it seems they wouldn't represent actual subjective interest.

To see the implausibility of the idea that agents who don't display integrated preferences in their choice behaviours nevertheless typically have actual mental states corresponding to the integrated preferences ascribed to them by the behavioural welfare economist, consider the fact that this would imply that any behavioural anomaly is a case of acting against one's own actual better judgement. As Sugden also highlights (p.80), it is implausible that this should be so regarding all behavioural anomalies. For instance, agents who violate expected utility theory by displaying Allais preferences aren't plausibly described as acting against some actual better judgement, even if we accept they exhibit some kind of irrationality.[4] I will take for granted that the resolution of the apparent dilemma just described does not come from resolving the first horn.

According to the second horn, if we take latent preferences to be merely hypothetical preferences, which don't need to directly correspond to actual mental states, they seem to be not subjective in the way they would need to be to satisfy the anti-paternalist ideal. If these preferences simply express what I would want under different circumstances, they

[4]More strongly in support of the first horn, I have argued elsewhere (Thoma forthcoming) that even agents whose choice behaviour can be captured with an integrated preference representation aren't in general plausibly ascribed mental states corresponding to these preferences, as such mental state ascription typically assumes too much cognitive processing, and adds little explanatory value.

may be wholly uninformative about what I actually currently judge to be in my best interests. And the anti-paternalist aims to respect an agent's own subjective judgements or choices. Suppose, for instance, that were I to be fully informed of all publicly available information, I would have no more desire to read and learn, and would always prefer other activities. This hypothetical preference tells us nothing about what is in my current, ignorant subjective interest regarding learning activities.[5]

This problem can potentially be avoided if we develop a concept of hypothetical latent preference that nevertheless preserves a link to our current subjective interests. And, to guarantee the normative authority of the latent hypothetical preferences, this link should be correct or faultless in some sense. Most plausibly, we could think of latent preferences as preferences one would arrive at through an ideal, well-informed reasoning process that starts from one's current subjective interests, and, to avoid the problem I just mentioned, does not alter one's fundamental subjective interests. What the proponents of the New Consensus would need to show, to keep their anti-paternalist credentials, is that such an ideal reasoning process would typically take agents whose choice behaviour is anomalous from their existing attitudes to the integrated preferences the behavioural economist ascribes to them.

Much of Sugden's case against the New Consensus in fact engages with the question of whether there is such a 'latent reasoning' (p.62) process that would produce integrated preferences. Before addressing this case, however, I want to highlight in the next section how this way out of the dilemma requires us to rethink the normative foundations of welfare economics, and specifically to abandon the idea that welfare measurement is *normatively* preference-based (in addition to formally representing welfare with a preference relation). This will serve as the basis of my defence of Bernheim and Rangel's (2007, 2009) choice-theoretic approach.

4 Subjective Interests Beyond Preference

To escape the second horn of the dilemma described in the last section, we need to develop a notion of hypothetical latent preference that preserves the right kind of link to an agent's actual subjective interests. This requires, I want to show here, a substantial rethinking of the way in which welfare economists conceive of the relationship between welfare, preference, and subjective interest, in at least three related senses. The last of these is an explicit abandonment of the idea that welfare measurement should be normatively preference-based.

[5]Note that this is also a standard criticism of 'informed desire' accounts of wellbeing. See Crisp (2017).

First, preferences are traditionally viewed as primitives in economics, and the process of preference formation is treated as a black box, as Sugden also remarks (p.63). The proposed way out of the dilemma, on the other hand, requires us to think of preferences as the result of a reasoning process that starts from underlying, more fundamental attitudes, and which, in the cases of interest to the authors of the New Consensus, has gone wrong. This view of preference is in fact explicitly accepted by at least one of the authors Sugden counts as part of the New Consensus, namely Bernheim (2016, pp.19-20), who concurs with a view popular among behavioural scientists (see, e.g., Lichtenstein and Slovic's 2006 collection *The Construction of Preference*) that preference is 'constructed' in the sense that we form overall attitudes to the options available to us only when called upon to do so, e.g. in order to make a choice.[6]

According to Bernheim, this construction takes place on the basis of underlying attitudes that have as their object various different attributes of the options open to the agent. The preference construction that agents engage in when called upon to choose proceeds by aggregating these various attitudes into an overall evaluation. For instance, customers in a cafeteria will have various desires of different strengths relating to the attributes of the snacks on offer, e.g. their healthiness, cost, taste, or novelty. Customers don't typically come equipped with a preference-like mental state that already ranks all the available options. But choice can be seen as involving the construction of such a preference, by aggregating the various prior attitudes the customer has to the attributes of the snacks on offer.

The second way in which getting out of the dilemma described in the last section requires a rethinking of traditional welfare economics is that it must involve making a distinction between two kinds of conative attitudes: those, on the one hand, that are the starting point of deliberation, and that shouldn't be changed by the reasoning process, and those, on the other hand, that may be formed in deliberation, and that, if there has been a mistake in reasoning, can be described as mistaken by the agent's own lights (as judged against her more fundamental attitudes). This distinction roughly maps on to the traditional distinction between attitudes that pick out ends and attitudes that pick out means. Again, Bernheim (2016, p.17) explicitly makes such a distinction between what he calls 'direct' judgements (about ends) and 'indirect' judgements (about means). Viewed through this distinction, preferences as formalized in economics, being intimately connected to choice, will typically correspond to attitudes about means, and capture which

[6]It might seem that accepting this view is problematic for behavioural welfare economists. According to Rizzo and Whitman (2020), who present a similar criticism of behavioural welfare economics as Sugden, it 'robs them of the Archimedean point that they would use to judge outcomes. If preferences do not exist independently of the act of choice, then there is no preference set against which to judge the individual's choices as deficient.' (p.58) This, however, assumes that preferences are the only attitudes that can potentially serve as a subjective standard against which to judge an agent's choices. But this is precisely what the rethinking of the normative foundation of welfare economics described here denies.

outcomes or options the agent takes to better serve her ends on balance. In the cafeteria example, we can think of customers having various ends that different snacks might serve, for instance, living healthily, tasting delicious things, and so on. A preference relation that provides an overall ranking of snacks can be seen as expressing judgements about how well the different snacks serve those ends on balance. And such judgements about means can in principle be mistaken by the agent's own lights.

The third aspect in which the proposed way out of the dilemma described in the last section involves a rethinking of traditional welfare economics is that it no longer views preferences (as formalised in economics) as direct representations of the attitudes that constitute subjective interest; i.e. it involves abandoning the idea that welfare measurement is normatively preference-based. Instead, subjective interest is now understood in terms of the fundamental attitudes that stand at the beginning of the reasoning process that produces preferences. Preferences and choice may or may not accurately capture subjective interest thus understood, depending on whether there has been a mistake in the reasoning process.

Where we have no reason to think that there has been a mistake in reasoning, we can of course still use revealed preference as a way of measuring welfare in practice, even if a preference-like mental attitude is not what constitutes subjective interest. The approach may also potentially still use a preference relation to formally represent welfare. At the same time, this approach does open the door, even for those with generally anti-paternalist inclinations, to no longer defer to revealed preference in cases where we have both good evidence that there has been a mistake in reasoning, and we know how to correct it, and to instead defer to the preferences the agent ideally would have formed on the basis of her ends. This amounts to condoning what is sometimes called 'mere' means paternalism, that is, paternalism that helps agents achieve their subjective ends, in situations where they are likely to choose the wrong means to their ends.

Again, Bernheim (2016) implicitly abandons the idea that welfare measurement should be normatively preference-based in just this way when he uses the distinction between direct and indirect judgements to justify the first step of his framework. This first step allows us to throw some choice data out as clearly based on a mistake. He proposes two criteria for classifying a choice as a mistake. First, it must arise from what he calls 'characterization failure', whereby a choice is 'predicated on a characterization of the available options and the outcomes they imply that is inconsistent with the information available to the decision maker' (p.48). And secondly, another available option would have been chosen were it not for the characterization failure (so the agent didn't choose the right thing for the wrong reasons). The idea here is that if these conditions are fulfilled, we can infer that the agent made a mistaken judgement about the best means to her ends. And, if it is only direct judgements – the judgements about ultimate ends that stand

at the beginning of deliberation – that the welfare economist ought to be deferential to, mistaken judgements about mere means may in principle be overridden and in any case do not track subjective interest. What defines subjective interest are the direct judgements about ends that stand at the beginning of deliberation, not preferences, which stand at the end.

If we want to avoid the dilemma described in the last section, I think we must in just this way reject the idea that welfare measurement is preference-based in the normative sense. If, as the authors of the New Consensus want to allow for, an agent's actual revealed preferences are sometimes mistaken while some other hypothetical preferences are correct, they must be mistaken or correct by some standard other than the agent's actual preferences. To satisfy the anti-paternalist ideal, this must be a subjective standard nevertheless, some set of conative attitudes of the agent's other than preference, which preferences are ideally correctly responsive to. If I have formed mistaken preferences, I have to go back to this standard, and reconstruct what preferences would result from an ideal reasoning process based on these more fundamental attitudes, rather than the ones I actually formed based on mistaken reasoning. If we were, instead, to insist on normatively preference-based welfare measurement, the New Consensus could be dismissed fairly quickly on the basis of the dilemma described in the last section: The notion that agents who choose anomalously have latent integrated preferences as actual mental states is psychologically implausible, and merely hypothetical preferences can't be shown to have the right connection to our actual subjective interests while holding on to a preference-based account of subjective interest.

What the rethinking of the normative foundations of welfare economics described in this section amounts to is that we shouldn't take preference-like mental attitudes, which are all-things-considered rankings of the options on offer, to constitute an agent's subjective interests. Rather, subjective interests are constituted by more fundamental attitudes, call them desires, on the basis of which preferences are typically formed. They define the ends agents ultimately aim to achieve by picking one option rather than another. And they typically have as their object various attributes of the options on offer, such as the tastiness, healthiness, and cost of snacks. Call this new normative foundation for welfare measurement 'desire-based'. It provides a way out of the dilemma of the last section in theory: Ultimately, we need a welfare measure that provides a ranking of options (even if only a partial one). On this picture, the welfare measure should capture how well different options fulfil an agent's underlying desires all-things-considered, which is something a hypothetical ideal reasoning process would reveal. Agents themselves may, for various reasons, sometimes fail to choose in line with what best fulfils their desires all-things-considered. In those cases, the true welfare measure would neither fail to be subjective (it would capture what serves the agent's actual underlying desires best) nor assume anything unrealistic (that the agent actually formed correct judgements but acted

against them for some reason).

But this solution in theory comes with a big challenge in practice: How can we know what serves people's ends best all-things-considered? And must there always be a determinate answer to that question? Agents themselves appear to aggregate their various desires when deliberating and choosing; however, we are now acknowledging that they can sometimes get this wrong. So how should the behavioural welfare economist proceed? This is a significant challenge. The next section will return to the core approaches within the New Consensus, and argue, in agreement with parts of Sugden's analysis, that in virtue of their use of an integrated complete preference relation, they do not respond to this challenge appropriately. However, the following section will make the case that Bernheim and Rangel's choice-theoretic framework can live up to the anti-paternalist ideal if interpreted in the normatively desire-based way I have outlined here.

5 The Case Against the New Consensus

Sugden characterises the New Consensus as committed to using integrated latent preferences as a measure of welfare. For this to be in line with the anti-paternalist ideal within the normatively desire-based framework just developed, it would, firstly, need to be the case that there is an ideal reasoning process that would lead any agent exhibiting anomalous behaviour from their current subjective interests to integrated preferences. Not only that, but behavioural welfare economists would also, secondly, need to have a way, in practice, to uncover what the integrated preferences are that capture what in fact serves the agent's subjective interests best on balance, even when her actual choices don't track this. We can understand some of Sugden's remarks on the failure of the model of the 'inner rational agent', and part of his discussion of the choices of SuperReasoner, as, respectively, an empirical and a normative challenge to these presumptions. In discussing his arguments, I will focus here on those authors of the New Consensus who appeal to a *complete* latent preference relation, and leave discussion of Bernheim and Rangel's approach to the next section.

The case of context-dependent consumer choice in the cafeteria serves as a useful illustration for both of Sugden's lines of argument. As previously mentioned, those in the New Consensus advocating means paternalist measures often focus on cases where consumer choice is influenced by seemingly irrelevant contextual factors, such as where on the shelf an item is located – in the case discussed at length by Sunstein and Thaler, consumers are more likely to pick an option that is situated at eye level – or whether the consumer is feeling hungry at the time of choice (even though choosing for a future time) – consumers are more likely to choose a nutritionally richer option when choosing while

hungry (see Read and van Leeuwen 1998). Sugden argues that such context-dependence is best explained by the psychology of attention: The change in context focuses our attention on some features of the options at the expense of others, leading us to give greater weight to them when making our ultimate decision (pp.68-69).

Choice in these cases is inconsistent and thus does not reveal integrated preferences. But those advocating for the New Consensus need to nevertheless think that we can ascribe integrated latent preferences to the agents in these scenarios, in relation to which at least one of the set of inconsistent choices will have been counter-preferential. But this, Sugden argues, we cannot do. His empirical criticism starts with a description of the 'inner rational agent' that does not assume her to actually have integrated preferences, and is thus at least prima facie more plausible: We can think of her as having an 'assumed capacity' (p.65) for exhibiting integrated preferences, but the reasoning process that would result in these integrated preferences remains latent as it is intercepted by some cognitive bias. Sugden argues this claim, too, however, is psychologically unrealistic (pp.67-72). The argument is, roughly, that even behaviour that reveals integrated preferences is in need of psychological explanation. And when we look at our best psychological theories, the assumed capacity to form integrated preferences plays no non-redundant explanatory role. The psychology of attention, for instance, explains each of a set of inconsistent choices in the same way, and thus accords no role to such a capacity.

Turning to the normative criticism, Sugden also questions in principle why even a perfectly rational agent, such as SuperReasoner, would need to display context-independent preferences. Sugden appears to give two main justifications for this idea. For one, he takes the psychology of attention that explains context-dependence to be driven by 'feelings' which he effectively treats as arational. These feelings would thus be shared by SuperReasoner and her ordinary human counterpart, and so the process leading to context-dependence is a dynamic SuperReasoner would also be subject to (pp.73-74).

Secondly, Sugden claims that context-dependence often occurs in situations where agents face a multi-attribute decision problem, requiring them to trade off several different features they might care about in their options, e.g., taste, healthiness, and cost. Note that, by appealing to a process of preference formation on the basis of underlying direct desires regarding attributes of the options on offer, Sugden here explicitly echoes the normatively desire-based approach as outlined in the previous section. What produces context-dependence, within this picture, is that insofar as our attitudes to those different features of the options are concerned, it may be simply indeterminate which of a number of options an agent ought to choose. Suppose you have to choose between a healthy, tasty but expensive option, a tasty, cheap, but unhealthy option, and a cheap, healthy, but bland option. You like snacks to be cheap, healthy and tasty, but your desires for these things may both be imprecise in their strength and different in kind, resulting in no one

uniquely correct way of trading them off. Call the claim that such indeterminacy exists the thesis of *non-uniqueness* of rational preference: an agent's underlying desires may not uniquely determine a rational integrated preference relation capturing what serves the agent's desires best on balance. It might seem to follow from non-uniqueness of rational preference that there is also nothing rationally problematic about context-dependent preference, that is, about effectively letting context decide how to resolve the indeterminacy (pp.74-75), e.g. choosing whichever of the three snack just described is presented at eye level.

I take the non-uniqueness thesis about rational preference to be highly plausible, given the multitude, vagueness and multifariousness of the desires relevant to many of the decision problems we face. However, it doesn't in fact follow from the non-uniqueness thesis that context-dependence of preference is rationally permissible. It might be, for instance, that while different ways of resolving indeterminacy are rationally permissible, rational agents should settle – arbitrarily – on one way of resolving the indeterminacy and then stick to it, in this way effectively displaying integrated preferences in their choices despite non-uniqueness. Pragmatic arguments grounded in the possibility of being exploited and making a sure loss unless one displays integrated preferences might establish a rational requirement to in this way display integrated preferences in one's choices despite non-uniqueness.[7]

If one is persuaded by such pragmatic arguments, one might thus question Sugden's claim that SuperReasoner, being perfectly rational, would display context-dependent preferences. However, as Sugden also notes in discussing a similar rejoinder (pp.75-77), this does not actually help the New Consensus. For one, if the integrated preference relation SuperReasoner would settle on is truly arbitrary, then the welfare economist presumably can't determinately reconstruct SuperReasoner's preferences, and the New Consensus approach would be unworkable. Moreover, even if we could determine which integrated preference relation SuperReasoner would arbitrarily settle on to resolve an indeterminacy, this does not resolve the indeterminacy for the actual human agent: Her subjective interests are not precise enough to be fully and accurately represented by a unique integrated preference relation. By appealing to a unique but arbitrary integrated preference relation as a welfare standard for means paternalist interventions, we may thus be imposing precision where there is none, and end up intervening where what the agent would have chosen herself is not actually against her interests. This does not live up to the anti-paternalist ideal.

These are problems that arise *in principle*: A normatively desire-based welfare measure should capture how well different option serve an agent's desires on balance. And

[7]See, however, critical discussion of such pragmatic arguments in, e.g. Cubitt and Sugden (2001), and, within the desire-based framework sketched here, Thoma (2017).

arguably, real agents' desires are often such that they don't allow for options to be ranked determinately. A welfare measure in the form of a complete integrated preference relation is thus inappropriate in principle, before we even consider the epistemic challenges in practice involved in reconstructing an integrated preference relation for agents that exhibit choice 'anomalies'. Even granting that there is a unique and complete integrated preference relation that captures what best serves an agent's ends on balance, upon observing context-dependent choice, it may be difficult to know which of the set of context-dependent choices do and which don't correspond to what is in the agent's interest on balance. For instance, how could we know whether it is really the healthy, tasty but expensive option, or the tasty, cheap, but unhealthy option that best serves the cafeteria customer's interests on balance after having observed her choose whichever is at eye level? We would need clear evidence that one of these choices was a mistake in light of the underlying desires for the feature of those options. But such evidence is hard to come by, short of talking to the agent.

Any approach that features a unique and complete integrated preference relation ascribed to agents who exhibit behavioural anomalies will definitively identify which of the agent's context-dependent choices are mistakes, and which are not. The worry is that this will often involve a big epistemic leap. My Thoma (2021) discusses how this is so even for the most sophisticated reconstructive approaches based on cumulative prospect theory. At the same time, it is in the spirit of the anti-paternalist ideal that the burden of proof for non-deference to an agent's own choices should be high. So there is both a problem in principle, and a epistemic problem with behavioural welfare economists ascribing unique and complete integrated preference relations to agents: This both presumes that agents' desires are more precise and can be more precisely aggregated than they often are, and oversteps the evidence usually available on which choices are mistakes and which aren't.

6 In Defence of the Choice-Theoretic Approach

The case against those approaches within the New Consensus that use a complete integrated preference relation as a formal welfare measure thus appears conclusive: Those approaches can be dismissed fairly quickly if we take them to also be normatively preference-based, that is, if we take preferences to also constitute subjective interest. And, as the last section showed, these approaches also do not hold up when the anti-paternalist ideal is implemented in a desire-based way. I here want to show, however, that on the desire-based picture, Bernheim and Rangel's choice-theoretic framework in fact comes out as a plausible approach to welfare measurement in line with the anti-paternalist ideal. In particular, the framework withstands the arguments discussed in the previous section, given the way it accommodates context-dependent choice.

As mentioned above, Sugden interprets Bernheim and Rangel's use of 'unambiguous choice' as a welfare criterion as an appeal to a notion of latent preference that, while being incomplete, is otherwise integrated and plays a role similar to the latent preferences featuring in the other approaches of the New Consensus. I take Sugden's key challenge to be this: If choice is context-dependent, why should our welfare measure be context-independent? This challenge is indeed compelling if we interpret Bernheim and Rangel's approach as normatively preference-based in the following way: We judge an agent to 'truly prefer' something only if she unambiguously chooses it. We then take the resulting coherent but incomplete 'true' preference relation to constitute the agent's subjective interests. For this to be a plausible approach, we would need to somehow explain why context-independent choices are taken to reveal true preference and thus to be welfare-relevant, whereas context-dependent ones are not. Here it seems the response would need to invoke some form of Sugden's 'inner rational agent': Only the context-independent choices are a true expression of rational agency. And then Sugden's previously discussed empirical and normative criticisms of the New Consensus become relevant again.

But, as we have seen in Section 4, Bernheim's (2016) more recent interpretation of the framework is not normatively preference-based, but in line with the desire-based account of what constitutes subjective interest I presented there. And, as I will now show, Sugden's challenge can be answered if we adopt a desire-based view of the foundation of anti-paternalist welfare measurement – which, as we have seen, is a possibility Sugden himself entertains in his discussion of the New Consensus. In fact, this interpretation neatly accommodates Sugden's own analysis of context-dependent choice.

On the desire-based account of what constitutes subjective interest, we can interpret Bernheim and Rangel's choice-theoretic framework as follows: Generally speaking, unless we have clear evidence, with a high burden of proof, of a mistake due to characterization failure, we should presume that actual choice is a good guide to what serves an agent's desires best all-things-considered. And so we should defer to those choices. What are we to make of context-dependent choice that cannot be explained by characterization failure? Here we should not assume that any of the set of inconsistent choices was mistaken. Rather, we should take each choice to represent a presumptively permissible way of aggregating the agent's desires. And so we should treat it as potentially indeterminate which of two options regarding which choices are context-dependent has higher welfare for an agent.

The indeterminacy in the welfare measure may have two legitimate sources: Firstly, the agent's underlying desires and their respective importance may simply not be precise enough to determine one unique and complete integrated preference relation that correctly aggregates them. This is an application of the non-uniqueness thesis which is, as we have seen, entertained by Sugden himself. It is also consistent with the attention-based

psychological explanation Sugden gives for context-dependent choice, as context could be what makes agents resolve the indeterminacy in one way rather than another. The other legitimate source of indeterminacy in our welfare measure is the high burden of proof for when a choice is a mistake, combined with a common paucity of evidence on the issue. The anti-paternalist should view it as better to presume it to be indeterminate whether it is really the healthy, tasty but expensive option, or the tasty, cheap, but unhealthy option that best serves the cafeteria customer's interests than to risk overriding a choice as mistaken on sparse evidence. If there is no clear evidence of mistake in either of the customer's context-dependent choices, I think this is as it should be.[8]

What makes unambiguous choice a good welfare criterion, on this interpretation, is that it appears to reveal that an option is unambiguously better than another in terms of the agent's underlying desires. Where choice is inconsistent, on the other hand, we simply live with this indeterminacy, resulting in an incomplete welfare measure. Importantly for us, this interpretation of the choice-theoretic framework withstands Sugden's critique. First, it does not rely on the idea that agents have integrated latent preferences as actual mental states, or even consistent but incomplete actual preferences: Bernheim (2016) explicitly denies that we should take unambiguous choice to reveal a context-independent 'true' preference (p.20). Even where choice ends up consistent, agents may still be aggregating only on the spot. And secondly, this interpretation also does not presuppose that ideally rational counterparts of actual agents would choose consistently. It may accept that, where there are different equally permissible ways of aggregating underlying direct desires, ideal rationality does not rule out picking one on the basis of arbitrary contextual variation. At the same time, the account is consistent with a view requiring ideally rational agents to choose in a way that reveals integrated preferences, e.g. due to the pragmatic arguments mentioned at the end of the last section. It would merely maintain that, to the extent that the preference relation settled on is arbitrary in light of the vagueness of the underlying desires, it does not reveal determinate welfare rankings for the actual human counterpart. Either way, to explain why unambiguous, context-independent choice reveals determinate welfare rankings, and context-dependent choice does not, we need not appeal to irrationality or defects in preference formation in the latter case. Rather, the difference can be explained by what these respective choice patterns reveal about the precision and force of the relevant desires they are based on.

[8]There is a worry here about whether the incompleteness of the welfare measure makes the approach practically unhelpful more generally. This is partly addressed by Bernheim (2016). Suffice it to say here that at least prima facie, it is not clear that the Opportunity Criterion as defended by Sugden is more helpful, given it also only produces incomplete rankings.

7 Conclusion

Traditionally, welfare economists have viewed integrated preferences not only as a formal measure of welfare, but also as direct representations of the mental attitudes that constitute subjective interest, and thus serve as the normative foundation of anti-paternalist welfare measurement. I have argued here that if we hold on to this account of subjective interest, the 'New Consensus' in behavioural welfare economics can be dismissed fairly quickly as either based on psychologically unrealistic assumptions or as not sufficiently subjective to live up to the anti-paternalist ideal. I then described a shift in thinking about the normative foundations of anti-paternalist welfare measurement that remains more or less implicit in the behavioural welfare economics literature, to taking underlying, more fundamental desires regarding features of the available options to be the conative attitudes that constitute subjective interest. Much of Sugden's critique of the New Consensus can be viewed within this desire-based framework. However, I have argued that Bernheim and Rangel's choice-theoretic framework withstands his critique if interpreted in a desire-based way. This framework takes unambiguous choice to reveal determinate welfare rankings while at the same time accommodating context-dependent choice, all without appealing to a problematic notion of an 'inner rational agent' to explain the difference.

To motivate a radically different welfare measure that gives up on trying to rank the objects of choice at all, as Sugden's opportunity-based criteria do, Sugden's critique of the New Consensus is thus not sufficient. This is not to say that the approach I have defended here can't be dismissed on another basis. One important worry appears to me to be this: While, in line with the anti-paternalist ideal as I have described it, the desire-based approach identifies a truly subjective basis for welfare measurement, it does condone what we called 'means paternalism'. Intuitively, this might be a less problematic kind of paternalism. But it is a type of paternalism nevertheless. Whether it is acceptable depends on the details of how one justifies the anti-paternalist ideal. Authors of the New Consensus sometimes write as if the deference to choice in the orthodox approach to welfare economics was only ever plausible under the assumption that agents are rational. But on the liberal justifications of the anti-paternalist ideal, at least, it's not entirely clear why that should be so. If we are happy to defer to individuals' judgements even when they are wrong about what is ultimately good for them, why shouldn't we be equally happy to defer to them when it comes to mistaken judgements about the means to fulfilling their ends?

Acknowledgments

I thank Douglas Bernheim, Måns Abrahamson, Constanze Binder and an anonymous referee for very helpful feedback on earlier drafts of this paper. I am also grateful to audiences at the LSE Workshop on the Normative Implications of Behavioural Economics in 2019, and the British Society for Philosophy of Science Annual Conference in 2021 for very useful discussion.

Biographical Note

Johanna Thoma is Associate Professor in the Department of Philosophy, Logic and Scientific Method at the London School of Economics and Political Science. Her research focuses on the philosophy of economics, decision theory, ethics and public policy.

References

Douglas Bernheim. The good, the bad, and the ugly: A unified approach to behavioural welfare economics. *Journal of Benefit-Cost Analysis*, 7(1):12–68, 2016.

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

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

John Bershears, James J. Choi, David Laibson, and Brigitte C. Madrian. How are preferences revealed? *Journal of Public Economics*, 92:1787–1794, 2008.

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

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

Roger Crisp. Well-being. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition, 2017.

Robin P. Cubitt and Robert Sugden. On money pumps. *Games and Economic Behaviour*, 37(1):121–160s, 2001.

Daniel Haybron and Anna Alexandrova. Paternalism in economics. In Christian Coons and Michael Weber, editors, *Paternalism: Theory and Practice*, pages 157–177. Cambridge University Press, 2013.

Gerardo Infante, Guilhem Lecouteux, and Robert Sugden. Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, 23(1):1–25, 2016.

Botond Köszegi and Matthew Rabin. Mistakes in choice-based welfare analysis. *American Economic Review*, 97(2):477–481, 2007.

Julian Le Grand and Bill New. *Government Paternalism: Nanny State or Helpful Friend?* Princeton University Press, 2015.

Sarah Lichtenstein and Paul Slovic, editors. *The Construction of Preference*. Cambridge University Press, 2006.

Paola Manzini and Marco Mariotti. Categorize then choose: boundedly rational choice and welfare. *Journal of the European Economic Association*, 10(5):1141–1165, 2012.

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

Mario J. Rizzo and Glen Whitman. *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*. Cambridge University Press, 2020.

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

Robert Sugden. *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford University Press, 2018.

Cass R. Sunstein and Richard Thaler. Libertarian paternalism is not an oxymoron. *University of Chicago Law Review*, 70:1159–202, 2003a.

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

Richard Thaler and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth and Happiness*. Yale University Press, 2008.

Johanna Thoma. *Advice for the Steady: Decision Theory and the Requirements of Instrumental Rationality*. PhD thesis, University of Toronto, 2017.

Johanna Thoma. Merely means paternalist? Prospect theory and 'debiased' welfare analysis. Unpublished manuscript, April 2021.

Johanna Thoma. Folk psychology and the interpretation of decision theory. *Ergo*, forthcoming.