Merely Means Paternalist? Prospect Theory and ‘Debiased’ Welfare Analysis

May 4, 2022

1 Introduction

Opposition to paternalism has a long tradition in economics: When designing and managing institutions, and when making public policy, we should generally respect people’s own judgements and choices about how they want to live their lives; we should not interfere in people’s affairs for their own good, at least not without doing our best to defer to their judgements; when economics is normative or prescriptive, it merely aims to help ensure efficient pursuit of the values and goals of the affected parties. The findings of behavioural economics, documenting various systematic deviations from the standard economic conception of ideal rationality, have only partially loosened economists’ anti-paternalist convictions. While many behavioural welfare economists now embrace at least some paternalist measures, they generally accept them only when these measures help agents pursue their own ends. That is, they accept only what I will call ‘means paternalism’: the kind of paternalism that respects people’s subjective non-instrumental values, that is, their ends, while helping them take the best means towards pursuing them, for instance by helping them overcome various biases in decision-making.[1] As a paradigmatic example, Thaler and Sunstein’s (2008) ‘libertarian paternalism’ aims to “make choosers better off, as judged by themselves” (p. 5).

Deviations from the standard economic theory of rationality are especially common in decision-making under risk, that is, when there is uncertainty about the outcomes of

[1] Dworkin (2019) calls this ‘weak paternalism’, but I will stick to ‘means paternalism’ in order to set it apart more clearly from ‘soft paternalism’ as advocated by Feinberg (1986). Soft paternalism consists in the interference with an agent for her own good in situations where her choices are essentially non-voluntary. Means paternalism, in contrast, may involve interfering with voluntary choices, as long as these stem from bad judgements about means. At the same time, for the means paternalist, the interference is constrained to helping agents achieve their own goals, and thus not necessarily what the paternalist thinks is best for her.

an agent's choices, but probabilities can be assigned to the various potential outcomes. Expected Utility Theory (EUT) is widely taken to be the correct normative theory of risky choice. According to this theory, agents are rational only if they can be represented as maximising the probability-weighted sum of utilities of the various outcomes. Violations of EUT are common, and Cumulative Prospect Theory (CPT), as proposed by Tversky and Kahneman (1992), is regarded as our best descriptively adequate theory by many. In CPT, outcomes are described as deviations from a reference point (i.e. normally the 'status quo'), and CPT can accommodate both loss aversion (roughly, the idea that losses against the reference point loom larger than gains), and probability-weighting (whereby probabilities are transformed before they are used to weight outcome-utilities).

If she accepts that EUT is the correct normative theory of choice under risk and CPT is our best descriptive theory of choice under risk, a means paternalist might strive to correct for people's deviations from EUT (so long as this can be done in a way that is not objectionably invasive). One tempting way to do so, proposed most prominently by Bleichrodt et al. (2001), but also defended, for instance, by Li et al. (2014) and Pinto-Prades and Abellan-Perpinan (2012), is to use CPT to measure an agent's utility function over outcomes, and then to use that same utility function in an expected utility calculation to determine what are then taken to be the normatively correct preferences over risky gambles for that particular agent.[2] This procedure has been described as a kind of 'debiasing' of people's preferences under risk, and I will refer to it as 'CPT debiasing' in the following. The hope is that it can inform policies designed to help agents achieve better health outcomes, or make better financial and insurance decisions.

This paper investigates whether CPT debiasing can be given a means paternalist justification. One reason why CPT debiasing is particularly attractive to means paternalists is that it appears to get around a common line of criticism of paternalist approaches in behavioural welfare economics. That criticism is that evidence of violations of ideal rationality usually comes in the form of evidence of inconsistency in preference or choice. For instance, libertarian paternalist policies are often targeted at situations where an agent's choices (e.g. whether to eat a healthy or an unhealthy snack) differ depending on irrelevant environmental factors. The problem in these situations is that there are usually multiple ways in which inconsistencies could be resolved. And the worry is that the would-be means paternalist has no way of telling which way of resolving the inconsistency is more authentically the agent's own (e.g. whether to consistently eat the healthy or the unhealthy snack). She either has no reliable way of determining an agent's underlying, and presumably rational 'true preference', or there is no such thing to begin with.[3] We

[2] Many others have proposed using CPT to obtain an 'unbiased' estimate of utility without explicitly endorsing plugging this utility into an EUT model. See Bleichrodt (2002), Bleichrodt et al. (2007), Booij and van de Kuilen (2007), Abdellaoui et al. (2007), Abdellaoui et al. (2008).

[3] See Sugden (2018), Infante et al. (2016), Whitman and Rizzo (2015) and Rizzo and Whitman (2020) on this line of criticism, and **?** for proponents of the idea of underlying 'true preference'.

sometimes might be able to determine what an agent would counterfactually judge and do if she were not subject to some bias. But it is not clear that the means paternalist can appeal to merely hypothetical judgements and choices as a welfare standard, as her goal is to help agents pursue their own actual ends.[4] In short, the worry is that in the face of behavioural anomalies, would-be means paternalists may not have a welfare standard that is both appropriately subjective and that yields definite recommendations to base means paternalist policies on.

The use of CPT in debiasing anomalous preferences under risk seems to help us respond to these worries in the domain of risky choice at least.[5] Using CPT, we can identify a utility function for agents who violate EUT. This utility function seems to capture some aspect of the agent's subjective value judgements. In particular, we might think it captures the agent's valuation of outcomes, and thus her ends. And so even though we judge her preferences over risky gambles — which we can interpret as judgements over means to those ends — to be irrational, we now seem to have a welfare standard that is subjective and therefore acceptable to the means paternalist. Moreover, by then plugging this measure of the agent's subjective valuation of outcomes into an EUT calculation, we can determine what the rational way is for the agent to pursue her own ends in contexts of risk.

The normative appeal of CPT debiasing depends, of course, both on EUT being actually normatively adequate, and on CPT being actually descriptively adequate. Both presuppositions have been challenged. Harrison and Ross (2017) and Harrison and Swarthout (forthcoming) argue that Rank-dependent Utility Theory (RDU) as developed by Quiggin (1982) actually has a better fit with choice data from well-designed laboratory experiments and that evidence in favour of CPT is at least inconclusive. EUT is moreover taken to be too restrictive as a normative theory by many, with some version of RDU, which features probability-weighting but not loss aversion, advocated as the correct normative theory by some (e.g. Buchak 2013). Others take neither loss aversion nor probability-weighting to offer grounds for paternalist intervention (e.g. Camerer et al. 2003).

I will nevertheless grant the normative adequacy of EUT, and the descriptive adequacy of CPT. My argument will, however, apply equally to debiasing using RDU.[6] What I will argue in this paper is that, even if we grant these presuppositions, CPT debiasing cannot be given a compelling means paternalist justification. Firstly, there are reasons to doubt that the utility function measured within a CPT framework provides us with a measure

[4] See Gruene-Yanoff (2012, 2018) on this problem for 'reconstructive' approaches in behavioural welfare economics.

[5] As noted, for instance, by Pinto-Prades and Abellan-Perpinan (2012).

[6] Krzysztofowicz and Koch (1989) propose a debiasing procedure for probability-weighting only. Wakker and Stiggelbout (1995) suggest using RDU to obtain an unbiased estimate of utility, but don't explicitly go as far as endorsing using this estimate in an EUT model.

that isolates the agent's evaluation of outcomes, or her ends. And secondly, even if it does, the resulting means paternalism is a problematic type of means paternalism that should be ruled out by the same considerations that motivate economists' opposition to ordinary paternalism. This is because CPT debiasing imposes risk neutrality in the pursuit of subjective non-instrumental value on agents. EUT does not imply such risk neutrality, nor is such risk neutrality a plausible requirement of instrumental rationality. Plausibly, risk neutral pursuit of one's ends is just one of the permissible ways of pursuing one's ends. In such contexts where instrumental rationality is permissive, those with anti-paternalist leanings should, as much as possible, defer to the agent's preferences regarding how to pursue her ends. And in that case, adjustments to CPT preferences, even though we grant they are irrational, will only rarely be permissible, and should be more minimal than those implied by CPT debiasing. I will end by outlining such a less interventionist approach to identifying potential means paternalist interventions in the context of risk.

2 CPT Debiasing

As descriptive theories, the aim of decision theories is to represent agents' preferences and/or choices with a convenient formalism that facilitates the prediction of yet unobserved choices. Insofar as the formalism plausibly captures aspects of an agent's underlying psychology, such as her beliefs and desires, we might take the theory to be explanatory as well. As normative theories, decision theories claim that agents rationally ought to have preferences and choose in such a way that they are well described by the decision theory.

In EUT under risk, agents' preferences over outcomes and lotteries (that is, probability distributions over outcomes) can be captured just in terms of the probabilities and a function assigning utilities to outcomes. Preferences then track the probability-weighted sum of the utilities of the possible outcomes of an agent's choices. Various representation theorems, most famously that due to von Neumann and Morgenstern (1944), henceforth vNM, show that preferences are representable as such if and only if they abide by a number of axioms, including transitivity and the independence axiom. Consequently, when EUT is accepted as a normative theory, the normative content of EUT is usually taken to be that agents ought to have preferences that abide by these axioms. Representability of a set of preferences within the EUT framework does not guarantee that utility captures, as a cardinal measure, some real psychological quantity the expectation of which is maximised, and indeed part of my argument below relies on scepticism regarding whether it does.

CPT, which incorporates elements both of original prospect theory (Kahneman and Tversky 1979) as well as RDU, includes significantly more structure than EUT.[7] First,

[7] I am loosely following the presentation in Koebberling and Wakker (2005) here.

outcomes are represented in terms of deviations from a reference point, usually the agent's status quo, which can potentially vary over a series of choices. Agents are then ascribed a basic utility function over these gain and loss outcomes. Next, we make room for one source of loss aversion (what Harrison and Ross (2017) call 'utility loss aversion') by defining a composite or total utility function, which coincides with basic utility for gains, but weights the basic utility for losses by a loss aversion parameter λ. Lastly, preferences over lotteries are not determined as a probability-weighted sum of outcome-utilities as in EUT, but rather as in RDU, allowing for the weighting of probabilities themselves: Probabilities are transformed by a weighting function, and outcomes ranked from best to worst. The total utility of the best outcome is multiplied by the weighted probability of getting at least that outcome. The total utility of receiving at least the second best outcome is multiplied by the weighted probability of receiving at least that outcome, minus the weighted probability of receiving at least the best outcome, and so on for the other outcomes. In the end we sum up, and preferences should track this sum. What the weighting of probabilities (henceforth 'probability-weighting') allows for is that agents can give proportionately higher or lower weight to better or worse outcomes than their probabilities. As Harrison and Ross (2017) note, apart from introducing other potential distortions from EUT, this introduces a second way of capturing loss aversion. Axiomatizations for CPT have since been developed, including under risk (see Chateauneuf and Wakker 1999), specifying strictly weaker conditions on preferences over lotteries than EUT for representability within CPT. CPT is therefore more permissive regarding what preferences over lotteries agents may have, at the same time as representing them with a more complex formalism.

We can now say more precisely what CPT debiasing involves. If CPT is descriptively adequate for some agent, then there will be some CPT model that has good fit with the agent's preferences (which are in turn inferred from her choices). From her choice behaviour, we can infer robust measures of the probability weights, λ, and, importantly, basic utility. We then take that basic utility function measured in our CPT model, and use it within an EUT model, to calculate the expected utility of the various lotteries the agent is choosing between. This way, we have eliminated both loss aversion and probability-weighting in the valuation of lotteries (although, interestingly, not other potential forms of reference-dependence). We conclude that the correct, unbiased preferences for this agent track this expected utility. And we consider policies that serve those corrected preferences.

One curious feature of this procedure should be noted straight away: Proponents of CPT debiasing assume that it is in principle possible to take utilities measured in one theoretical framework (CPT), and then use them in another (EUT). This only makes sense if we think that utilities are meant to represent the same thing in both theoretical frameworks. In fact, this is a common presupposition in the literature out of which the CPT debiasing proposal emerged: Based on the observation that applying EUT to attempt to measure utility results in inconsistent utility measurements, various authors have tried to find alternative measurement procedures, based on different decision theories, that would result in a consistent measurement of utility.[8] These authors clearly think of utility as a measure of some theory-independent quantity, and of different decision theories that feature utility functions as different potential tools for measuring it. Since utilities in both CPT and EUT are unique up to positive affine transformations, utility is a cardinal measure: ratios of utility differences are meaningful. Whatever utilities in both frameworks are taken to represent by behavioural welfare economists must thus be something that comes in degrees.

What is this theory-independent quantity that utility is intended to be a cardinal measure of, according to the behavioural welfare economists in question? Bleichrodt et al. (2001) frame the task of utility measurement in the context of helping a client make better decisions as aiming to get an accurate measure of "clients' values" (p.1500), and to "better represent the interests of the client." (p.1510) Similarly, Abdellaoui et al. (2008) express the ambition behind finding an unbiased utility measure to be to enable us to make decisions in a client's best interest. Writing in the context of health-related decision-making, Pinto-Prades and Abellan-Perpinan (2012) take utility to be "the value we associate to health-related quality of life" (p.573), and Oliver (2003) speaks of different decision theories as frameworks for eliciting "cardinal health state values" (p.659). What they appear to have in mind is that utility is meant to be a cardinal measure of the degree to which an agent subjectively values the potential outcomes of the decisions to be made.

Policies designed to promote CPT debiased preferences, if agents have not explicitly consented to them, are intuitively paternalist. As we noted above, economists are usually opposed to paternalism. However, proponents of these measures might defend them as 'merely means paternalist'. Such a defence would first assert that these measures are only paternalist regarding how an agent's ends should be pursued. They do not impose ends on agents. And secondly, such a defence would try to show that means paternalist policies are in fact less objectionable than other forms of paternalism. The next section will present the best case for such a means paternalist defence of CPT debiasing. It will consider typical motivations for anti-paternalism, and why means paternalism is usually less objectionable than other forms of paternalism. It then presents the case for thinking CPT debiasing is a form of 'mere means paternalism'.

[8] See Krzysztofowicz and Koch (1989), Wakker and Stiggelbout (1995), Bleichrodt (2002), Oliver (2003), Oliver (2005), Bleichrodt et al. (2007), Booij and van de Kuilen (2007), Abdellaoui et al. (2007), Abdellaoui et al. (2008).

3 The Means Paternalist Defence of CPT Debiasing

In the most general terms, paternalism is interference with a person's actions or affairs, without her consent, motivated or justified by her own good. More precise definitions often work by specifying the nature of the interferences that may count as paternalist, and the nature of the motivation and justification of the interference. While traditionally, paternalism has often been thought of as the restriction of an agent's liberty for her own good (e.g. in the form of legal bans on unhealthy products), in the context of welfare economics, much less invasive measures are often thought of as paternalistic. For instance, a welfare state that hands out in-kind benefits when it could have handed out monetary benefits to those in need is often thought of as paternalistic insofar as the measure is motivated by the recipients' own good, even though, when compared to a no-benefits world, the measure increases rather than decreases opportunities for choice.

This is as it should be if we think of the characteristic harm or wrong of paternalism as a lack of respect for an agent's own choices and judgements in matters where we should defer to the individual, and grant that personal consumption decisions should be under a benefit recipient's own control. I thus agree with Haybron and Alexandrova (2013) and Hausman (2018) that any effect on a person, even if it is liberty-preserving or liberty-enhancing, could potentially be paternalistic, insofar as it concerns only the person's own wellbeing, or matters that should be under the person's control or should fall under the person's judgement, while showing a non-deferential attitude to the agent's own judgements and choices.

Different paternalist policies may be more or less severe, and we can identify at least two dimensions along which they can be more or less severe: Paternalist policies can be more or less intrusive, with libertarian paternalist policies designed to be minimally intrusive; And they can exhibit a more or less non-deferential attitude to the agent's choices and judgements. I will not discuss the first dimension any more in the following, as it is along the second dimension that means paternalism is usually thought of as being less problematic. It remains understood that the intrusiveness of the policy should also be taken into account when making final judgements about particular cases. There is clearly additional harm in the restriction of an agent's freedom or in the use of physical force.

Why might one think that policy-makers and the economists advising them should respect people's choices or defer to people's judgements regarding their own well-being? There are at least four common justifications for this anti-paternalist conviction. The first appeals to a subjectivist conception of what well-being is. If we think that well-being just consists in the satisfaction of individuals' actual preferences, which is a common conception of wellbeing in economics, then deferring to people's judgements and choices is just what one should do in order to promote wellbeing. A second justification does not rely on such subjectivism about wellbeing, but claims that people generally are in a much better position to make accurate judgements about what is good for them than a policy-maker or economist.[9] A third justification is not welfarist, but appeals to a core principle of liberalism, namely liberal neutrality – the idea that the state or other authorities should not impose any particular conception of the good life on its citizens, but should rather remain neutral between competing conceptions, so as to accommodate the inevitable plurality of conceptions of the good life. And a fourth type of justification holds that there is a distinct and non-derivative harm or wrong involved in interfering for an agent's good without deferring to her judgement, e.g. because it is insulting, as argued by Quong (2011), or an impermissible intrusion into what is rightfully for the individual to decide, as argued by Shiffrin (2000), Groll (2012), and, more specifically in the context of welfare economics, Sugden (2018).

Whichever is our favourite justification for anti-paternalism, means paternalism appears to turn out less problematic than other forms of paternalism. This is because means paternalists do show a deferential attitude at least to agents' judgements about their ends or ultimate objectives, or what are variably called their direct value judgements (e.g. by Bernheim 2016), or judgements about non-instrumental, intrinsic, or final value. Means paternalists show a non-deferential attitude only regarding agents' judgements about or choices of means to their ends, or their indirect value judgements, or judgements about instrumental value. The means paternalist overrides people's instrumental judgements about means to their ends, in order to help them pursue their ends.

As means paternalism still involves overriding an agent's judgements that pertain to her own wellbeing (albeit indirect judgements), I think it still counts as a form of paternalism. But even those with generally anti-paternalist convictions might judge it to be justifiable when uninvasive, given it ultimately serves to help agents pursue their own values. Moreover, the four standard justifications for anti-paternalism just described don't apply to the same extent to means paternalism. First, the only plausible subjectivist accounts of wellbeing take wellbeing to be constituted by an agent's preferences regarding their ultimate ends only, and not by their preferences over means to the realisation of those ends. For instance, where an agent's preferences over means to her ends are tainted by false beliefs, we generally don't think satisfying her preferences over means always makes her better off.[10] Second, the claim that people are in a better position to make accurate judgements about what is good for them is likely to be true in a wider range of

[9] This is a key part of Mill's (1859) case for his Harm Principle, which is an anti-paternalist principle. Hausman and McPherson (2009) argue that treating preferences as evidence of wellbeing is the more promising way to defend the importance of preference satisfaction in welfare economics than adopting a subjectivist conception wellbeing.

[10] To bracket this additional potential motivation for means paternalism, in the below discussion we will assume that the policy-maker and the potential target of CPT debiasing have access to the same information.

circumstances for judgements about ultimate ends rather than instrumental judgements. Again, judgements about means that are tainted by false beliefs are a case in point: A policy-maker might, for instance, have more accurate beliefs about the likely health outcomes of some activities, and know to have more accurate beliefs. And lastly, means paternalists do seem to respect liberal neutrality, and show respect, where it counts most, namely regarding judgements about ultimate ends.

In addition to Thaler and Sunstein (2008), means paternalism is advocated by many if not most behavioural welfare economists, insofar as they would like to keep a door open for correcting for irrational judgements about means, as they seem to be revealed in findings of behavioural anomalies. For further explicit defences, see, e.g., Camerer et al. (2003), Bernheim (2016) and Le Grand and New (2015). Returning to our central topic, the question now is whether CPT debiasing can be given a means paternalist justification. I take it that at least two things need to be shown in order for some policy to be defensible as 'merely means paternalist' towards an agent: First, the policy-maker needs to have some reliable way of determining what the agent's relevant ends are. And second, the policy-maker needs to be confident that she can make a superior judgement about the best means to the agent's ends than the agent herself. If the first condition fails, then the means paternalist has no way of being deferential to the agent's judgements where it counts, so that the general anti-paternalist considerations count against the measure. And if the second condition fails, the policy loses its positive appeal of helping agents serve their ends better, in which case it seems we should err on the side of deference to the agent's judgements.[11]

Importantly for my argument later on, the second condition also rules out paternalist interventions in some cases where the policy-maker can both determine the agent's ends, and an effective way for the agent to pursue them, namely in situations where the policy-maker imposes one effective way of pursuing her ends on an agent in a situation where the agent merely pursued a different, but equally effective and rationally permissible way of achieving her ends. Take, for illustration, a situation where two roads (which have no intrinsic merits) lead an agent to her goal equally well. She would choose the left road if left to her own devices, but a policy-maker imposes the choice of the right road on her. The policy maker does respect her goals, and is proposing one effective way of pursuing them – just not a better way than the one the agent would have chosen herself. In such

[11] Paternalist policies will often affect more than one agent, of course, and it is often impossible to design policies such that they serve everybody's ends, or to even find out what would serve each person's ends. Such policies may still be justifiable on means paternalist grounds insofar as they help a significant subset of agents serve their ends well, and don't cause significant harm to others. The justification for overriding those other agents' judgements about means would then not be their own good (and thus would not be paternalist), but the good of those agents we are being means paternalist towards. See Parry (2017) for discussion of this type of case. In any case, the two conditions I describe here need to hold true for at least some agents in order for a policy to have a means paternalist justification.

situations, too, the policy seems to have no positive appeal of helping agents serve their ends better than they would themselves, leading to the conclusion that the policy-maker should err on the side of non-interference and deference. In fact, in situations where there is rational leeway in how to pursue one's ends, it seems a special kind of liberal neutrality might apply, demanding that the policy-maker should not only refrain from imposing a particular view of the good life, but also from imposing any particular one of the rationally permissible ways of pursuing one's idea of the good life.

As pointed out in the introduction, doubts are often raised about paternalist policies proposed by behavioural economists that point to the failure of one of the two conditions just spelled out. The mere observation of behavioural inconsistency points to no particular way of resolving that inconsistency that would honour the agent's ends, and do so better than the agent's own choices. CPT debiasing seems different as there is initial plausibility to both conditions holding. Regarding the first condition, we have just seen that in the literature where CPT debiasing is proposed, it is commonly held that the basic utility function identified within CPT provides us with a cardinal measure of an agent's subjective valuations of outcomes, or in other words her ends. And regarding the second condition, accepting the normative adequacy of EUT, as we have seen proponents of CPT debiasing also do, might seem to imply that maximising the expectation of such a cardinal measure of subjective outcome valuations is the only rational way of pursuing one's ends – in which case CPT debiasing interferes with incorrect instrumental judgements in order to replace them with the only correct ones, fulfilling the second condition for a policy-maker using CPT debiasing.

In the rest of this paper, I will raise doubts about both conditions in fact being met in the case of CPT debiasing. First, I will argue in the next section that there is no special reason to think that the basic utility function identified by CPT should provide us with a cardinal measure of an agent's ends, though ultimately this is at least partly an empirical question on which there is inconclusive evidence. And secondly, section 5 will argue that even if the CPT basic utility function provided us with a cardinal measure of the agent's ends, EUT does not imply that an agent must maximise the expectation of *that* utility function in order to be instrumentally rational. Instrumental rationality and EUT are more permissive than that. In the face of such permissiveness, the anti-paternalist must be more deferential to the agent's original preferences over lotteries, that is, her preferences regarding how to pursue her ends, than CPT debiasing.

4 The First Challenge: Isolating Ends

CPT debiasing can only be means paternalist if the basic utility function identified in CPT models in fact provides us with a reliable and complete measure of all the agent's relevant ends, that is, the degree to which she subjectively values the potential consequences of her actions. I here want to raise three worries about its ability to do so. First, note that the basic utility function can only be a reliable measure of the agent's ends if everything the agent subjectively values as an end can be reduced to a property of outcomes. In other words, agents cannot intrinsically care about irreducible features of lotteries. For instance, attitudes regarding the thrill of gambling, or the anxiety of uncertainty, or structural features of gambles such as their mean, mode and variance (as discussed by Lopes 1981, 1996) are not naturally described as attitudes to outcomes. Nevertheless, we might think that these attitudes represent non-instrumental subjective valuations: Such agents may care about gambles not merely as means to getting good outcomes, but rather see some features of gambles as ends in themselves. And if they do, a utility function over outcomes, such as the basic utility function in CPT, cannot capture all an agent's ends in the context of risk. [12]

Arguing along these lines, in the philosophical decision theory literature, Stefansson and Bradley (2015, 2019) have recently denied the idea that a clear distinction can be drawn between preferences over lotteries being merely instrumental (in the sense that lotteries are valued just as means to achieving good outcomes) and preferences over outcomes expressing non-instrumental subjective valuations, that is, the agent's ends. In the decision theory they develop, both preferences over lotteries and preferences over outcomes can express an agent's ends. If that is so, unless we know we are dealing with the special case of an agent who values lotteries only as means to good outcomes, we cannot dismiss an agent's preference over lotteries merely as a bad choice of means to her ends. And we cannot use the basic utility function identified in CPT as a complete representation of the agent's ends. If we were to use it as it is used in CPT debiasing, we would be disregarding some of the agent's relevant ends in risky contexts – those that are only expressed in her preferences over lotteries.

Of course, this picture is consistent with saying that CPT preferences are ultimately irrational. But the kind of irrationality involved in CPT preferences cannot, or at least need not be of the purely instrumental type, because preferences over lotteries should not be evaluated purely instrumentally, by whether they serve the agent's preferences over outcomes well. Instead, on this picture, rational restrictions on all types of preferences are

[12] Intrinsically valuing such features of lotteries is often taken to be irrational by economists. However, this verdict must be based on a more substantive notion of rationality, one that evaluates an agent's ends and not only her means to those ends. Paternalist interventions based on such a substantive notion of rationality would no longer be merely means paternalist.

better interpreted as restrictions on what combinations of ends an agent may have. If CPT preferences are irrational, it must then be because there is an incoherence in the agent's ends. This would be unfortunate, but it is also the kind of irrationality that is not suitable for means paternalist intervention. We would be faced again with the problem that there are multiple possible ways of resolving the incoherence, and we cannot pick one out as a better way of respecting the agent's subjective values. Resolving incoherence in an agent's ends in a non-deferential way amounts to ends paternalism, not means paternalism.

A common response to this in the philosophical literature is to insist that, insofar as attitudes that appear to be irreducibly about lotteries are not merely instrumental, we can in principle redescribe outcomes so that these attitudes can be captured by a utility function over outcomes after all. If needs be, we could even include in the description of outcomes the description of the gambles the outcome was part of. Buchak (2013) calls this 'global individuation', and it is defended by Pettigrew (2015). Even if it seems unnatural, we could simply treat it as a modelling norm that outcomes need to be described such that everything an agent values non-instrumentally is captured in the outcome description.[13] There is some evidence that this is indeed treated as a modelling norm by economists, as calls for redescription of outcomes is a common response to at least some behavioural anomalies (e.g. cooperation in the Prisoner's Dilemma, or gambling behaviour by otherwise risk averse agents).

At the same time, however, outcomes in most economic applications, including applications of CPT, are described in very simple terms, for instance as mere monetary gains and losses. Moreover, the redescription strategy is only a convincing response for the proponent of CPT debiasing if we have a reliable way, in practice, to determine how outcomes need to be described in order for the utilities measured within CPT to effectively capture all the agent's ends. And this is the second worry I want to raise in this section. It is a harder problem than merely finding a model that has a good enough fit with the choices we observe. Imagine two agents who make exactly the same choices, and are representable with the same CPT model, say with a basic utility function defined over simple monetary outcomes. It is entirely possible that one of them genuinely only cares about outcomes described in monetary terms, and sees lotteries merely as means to good monetary outcomes, while the other one genuinely cares about features of lotteries as ends in themselves, such as the probability of not making any loss. For descriptive purposes, this difference might be irrelevant (which is one reason why many economists aim to avoid committing themselves to particular psychological causes of choice behaviour). But it is not irrelevant for the prescriptive purposes of the means paternalist. CPT debiasers, for

[13] Another response, advocated by Buchak (2013) herself is to say that many attitudes to global features of lotteries really are merely instrumental attitudes, that is, preferences regarding *how* to pursue one's ends. I have some sympathy for this idea, but this move does not help CPT debiasing, as it either leads to the rejection of the normative adequacy of EUT (as in Buchak's own case), or reinforces my argument below regarding the permissiveness of instrumental rationality and of EUT.

the reasons we saw above, are committed to a particular psychological interpretation of what the utility function should capture: it must be a cardinal measure of an agent's ends. The problem this example illustrates is that choice behaviour alone may not allow us to distinguish between cases where it is such a measure and cases where it is not. The two agents make the same choices, are ascribed the same utility function in a CPT model, but that utility function does not capture all of the second agent's ends, while it does for the first. CPT debiasing might thus point to a legitimate means paternalist intervention for the first agent. For the second agent, it does not, or at least not using this CPT model. To distinguish between the two cases, we need further information about the agents' subjective values, beyond choice data. In practice, and especially for large scale applications, this will be very hard to come by.

A related practical issue arises and is discussed in the CPT literature when distinguishing the basic from the final or composite utility, which includes the loss aversion parameter λ as a weight on the basic utilities of losses relative to the reference points. Formally, utility loss aversion could equally well be described by a reference-point dependent kink in the basic utility function, as by a basic utility function being biased by a loss aversion parameter. Again both models could be used interchangeably for descriptive purposes. But the difference is crucial for the normative purposes of the means paternalist, as it changes the way in which we identify the agent's ends. And again, the difference can be determined only by knowing more about the agent's values: Does she genuinely care more about losses than about gains, or is it rather that the loss frame biases her towards giving more weight to losses than her true values warrant?[14] Harrison and Ross (2017) take utility loss aversion to at least frequently express a sentimental response to losses that should not be overridden by the policy-maker, while allowing for paternalism in correcting for probability-weighting. And note that even probability-weighting could be interpreted as being rooted in genuine differential valuation of outcomes depending on how comparably bad they are.

What emerges from this discussion is that even once we have found a CPT model with good empirical fit, the normative interpretation of its parameters, and in particular

[14] Proponents of CPT debiasing admit that this is a crucial question. Koebberling and Wakker (2005), for instance, consider potential genuine reasons for 'intrinsic loss aversion'. Bleichrodt et al. (2001) write: "Loss aversion designates, in this paper, a deviation from expected utility, depending on psychological perceptions of reference points sensitive to strategically irrelevant reframings of decisions. It is this loss aversion that generates discrepancies between probability- and certainty-equivalent measurements. If there are intrinsic reasons why losses with respect to a status quo are more serious than corresponding gains, then we consider this effect as part of the genuine von Neumann-Morgenstern utility function. It belongs to the expected utility model and does not depend on irrelevant reframings. Our correction proposal concerns only the former loss aversion (...)." (p.1500) The authors do suggest here that 'intrinsic' loss aversion could be distinguished from loss aversion as captured by λ as it is not dependent on 'irrelevant reframings'. However, practically parsing the two attitudes in this way will be difficult, and moreover, the assumption that an agent's intrinsic valuations cannot be frame dependent is a substantive restriction on what kinds of ends an agent may have and thus against the spirit of means paternalism.

the question of whether the basic utility function effectively captures the agent's ends, can remain controversial, or, to use McQuillin and Sugden's (2012) term, essentially contestable.[15] Of course, there are ways of resolving this controversy in particular cases. But this would require information about agents' values that is hard to infer from traditional economic data. We would need to know quite a bit about the agents' psychology. And while we might be able to enquire into an agent's subjective values in a laboratory setting, in the field, where ultimately we would like to implement means paternalist policies, the relevant information will be harder to come by. This problem is especially devastating given CPT debiasing was developed specifically for cases where preferences cannot be debiased directly in conversation with a client.

One tempting response here is to claim that as in the case of descriptive models, we can make various approximations and idealisations as long as the models serve their purposes well enough, or can be expected to describe most people well enough. However, for the means paternalist normative project, this response is not good enough. Recall that one prominent justification for opposition to ordinary paternalism was that agents generally are in a better position to know what is good for them, and policy-makers are often in a poor position to do so, in which case it seems better to err on the side of deference to the agent herself. Means paternalism was supposed to do better. Now indeed we are in a situation where the policy-maker is in one sense in a superior position to the agent, in that she may be confident that the agent is in violation of what we accept as the normatively correct theory of rational choice. But there is no point in overriding irrational preferences when we are not confident we can help the agent pursue her own goals any better. The contestability of the normative interpretation of CPT models alone should undermine any such confidence in most applications, as it is not clear the policy-maker can determine the agent's ends sufficiently well. Agents may not be serving their ends well, but in contrast to the policy-maker, at least they can be sure what their ends are.

A third worry remains even if we think that the outcomes the basic utility function in our CPT model ranges over capture everything the agent non-instrumentally cares about, and that neither loss aversion nor probability weighting are to any extent explained by the agent valuing features of lotteries non-instrumentally. Even in such a case, we can't be guaranteed that the basic utility function provides us with a *cardinal* measure of the agent's subjective valuation of outcomes. To see that, I'd first like to point out that the same isn't guaranteed within EUT either. For some function to be a cardinal measure of an agent's subjective valuation of outcomes, or her ends, not only must the function order outcomes according to how much the agent values them, but the shape of the function must also reflect the degree to which the agent subjectively values outcomes. For my argument, it will be useful to have a hypothetical example of which it is uncontroversial that an accurate cardinal measure of the agent's subjective valuations would be linear in

[15] This problem is also raised by Bernheim (2016).

some good of interest. So, this must be an example of an agent who subjectively values each unit of a good exactly the same. Take, for instance, the Cookie Monster.[16] To be ecumenical between different accounts of what it might mean to subjectively value an outcome, imagine that the Cookie Monster has an equally strong desire for any cookie, judges each cookie to be equally useful or good for him, is equally committed to the consumption of any cookie, gets the same amount of pleasure out of any cookie (and so on for any other notion of value one might have). We should say then, on any account of value, that he values each cookie to the same degree, no matter how many cookies he has already had, and that thus a true cardinal measure of his subjective valuations would be linear in cookies.

Now suppose Cookie Monster abides by the axioms of vNM expected utility theory. Must the utility function in an EUT model representing his preferences be linear in cookies? Clearly not, because Cookie Monster may be risk averse or risk loving. Within EUT, agents who are risk averse or risk loving with regard to some good must be assigned a utility function that exhibits decreasing or increasing marginal utility respectively with regard to that good. For instance, if, for any lottery over different amounts of cookies, the Cookie Monster prefers some sure amount of cookies below the expected number of cookies in the lottery, we must assign decreasing marginal utility of cookies to him. And we must do so despite him valuing each cookie to the same degree. The EUT utility function thus need not be a cardinal measure of an agent's subjective valuations, even if it ordinally tracks them.

For those interested in identifying a function that provides a true cardinal measure of subjective valuations, as the behavioural economists who are my target here are, the underlying issue is that capturing risk aversion with decreasing marginal utility in an EUT model blends together two kinds of psychological causes of risk aversion: One might be risk averse because the more one already has of a good, the less one values it. Or one might be risk averse because one simply prefers to err on the side of not taking chances, and of ensuring one ends up with decent outcomes rather than risking all for the chance of ending up with more. This second kind of risk aversion is often called 'pure' risk aversion (and respectively we speak of pure risk seeking attitudes), and I will follow this terminology here, meaning it to refer to any risk aversion that is not explained by an agent valuing a good less the more she already has of it.[17] Cookie Monster's risk aversion is pure: he values all cookies the same; nevertheless, he is risk averse. This must simply be because he does not want to risk ending up with fewer cookies. The important point from this for our discussion is that in the absence of pure attitudes to risk, the EUT utility function may well be a cardinal measure of subjective value. But in the presence of pure attitudes

[16] See also [redacted] on this example.

[17] See Buchak (2013) for extensive discussion of the distinction between these two kinds of psychological mechanisms.

to risk, such as in Cookie Monster's case, it clearly is not. The shape of the EUT utility function may, in addition to degrees of subjective value, also capture pure attitudes to risk.[18]

If the shape of the utility function in EUT can in part express pure attitudes to risk and need not be a cardinal measure of an agent's subjective valuations, it is not clear why the same shouldn't be true of the basic utility function in CPT. As in the case of EUT, merely abiding by the axioms that guarantee CPT representability does not imply that the utility function must be a cardinal measure of subjective value. For that to be the case, we would need to be sure that any pure attitudes to risk are fully captured by probability weighting and loss aversion. Confidence that this is so can ultimately only come from looking more closely at the psychology of choice. Here, there is in fact some cause for optimism for the proponent of CPT debiasing. Utility functions measured within EUT systematically diverge from utility functions determined in riskless contexts using strength-of-preference judgements, which we might argue are an adequate cardinal measure of subjective valuation of outcomes. But there are at least some studies, e.g. Abdellaoui et al. (2007) and Stalmeier and Bezembinder (1999), that suggest that CPT utilities and the riskless measures do coincide. Nevertheless, these studies are limited in the kinds of lotteries studied. For instance, Abdellaoui et al. (2007) consider only two-outcome monetary lotteries where both outcomes are gains (and thus do not distinguish between CPT and RDU). Given the relative sparsity of evidence, the claim that the utilities identified in a particular CPT model provide us with a cardinal measure of the agent's subjective valuations also seems to remain essentially contestable. Consequently, the means paternalist usually can't be confident enough she can identify a cardinal measure of the agent's ends to base an intervention on it.

5 The Second Challenge: Permissiveness of Instrumental Rationality

Even though the last section raised a number of doubts about this, let us grant now that the basic utility function identified in CPT models provides us with a cardinal measure of an agent's complete non-instrumental subjective valuations, that is, her ends. CPT debiasing now proceeds by plugging the utility function obtained in the CPT model into

[18] See Dyer and Sarin (1982) for an early paper exploring the distinction between a subjective value function which would express only strength of preference for outcomes, and vNM utility, which may also capture pure risk aversion. Adler (2019), Chapter 2 and Appendix D makes a parallel argument in the case of wellbeing measurement: The vNM axioms alone cannot guarantee that utility provides us with a cardinal measure of wellbeing. A further assumption, which he calls 'Bernoulli' and essentially calls for risk neutrality with regard to one's wellbeing, is needed. Without it, all we get is that vNM utility and Adler's preference-based wellbeing measure are increasing functions of each other.

an EUT model. For prescriptive purposes, the CPT debiaser recommends lotteries to an agent (or chooses them on her behalf) maximising the expectation of that basic utility obtained in the CPT model. This would be uncontroversially doing better than the agent herself does if maximising the expectation of the CPT-obtained utility, which we are now granting is a cardinal measure of subjective value, is the uniquely most effective way to serve the agent's ends. But in this section, I will argue that instrumental rationality and EUT are more permissive than that. I will then explain why this undermines CPT debiasing, even if the CPT utility function is a valid cardinal measure of the agent's ends.

I take it to be intuitively uncontroversial that instrumental rationality is permissive under risk in the following sense: Keeping fixed all an agent's subjective valuations of outcomes, instrumental rationality does not prescribe a unique preference relation over lotteries the agent must adopt in pursuit of good outcomes. An example should help to support this intuition. Take again the case of the Cookie Monster, who, on any notion of what it means to subjectively value outcomes, values all cookies the same. Now he is given the choice between 47 cookies for certain and a 50/50 chance between 0 or 100 cookies. Which should he choose? I submit that both answers are rationally permissible for the Cookie Monster. There is some rational leeway in how he may rationally choose to pursue his cookie-eating goals.

Suppose the Cookie Monster prefers the 47 cookies for certain. And through inquiring more into his preferences, we find out he reaches a point of indifference between the lottery and a sure outcome when the sure outcome is 45 cookies. Below that, he prefers the lottery. Now imagine the Cookie Monster has a cousin, Cookie Aficionado, who is like the Cookie Monster in her evaluation of outcomes: She values all cookies the same. But, unlike the Cookie Monster, she prefers the lottery to the 47 cookies for certain. For her, we find out, the point of indifference is at 50 cookies. By presupposition, Cookie Monster and Cookie Aficionado have exactly the same ends (even though, within EUT, we will have to assign them different utility functions). They merely pursue them differently. And, intuitively, neither of them is instrumentally irrational. If we grant this, then we grant that instrumental rationality is (within bounds – we may rule out extreme attitudes) permissive under risk.

Note that in the terms of our earlier discussion, Cookie Monster, but not Cookie Aficionado displays some pure risk aversion. So the claim that instrumental rationality is permissive under risk amounts to the claim that some non-neutral pure attitudes to risk are rationally permissible. As we noted in the previous section, EUT does not rule out non-nuetral pure risk attitudes, as long as agents abide by the vNM axioms. For the same reason, EUT is not incompatible with permissiveness under risk. Cookie Monster and Cookie Aficionado could both be expected utility maximisers. They would need to be represented with different utility functions: concave for the Cookie Monster and linear for

Cookie Aficionado. This is compatible with them nevertheless valuing cookie outcomes in the same way. We just need to accept that for at least one of them, Cookie Monster in this case, the utility function is not a cardinal measure of her ends.

Thus, I take it that intuitively, instrumental rationality is permissive under risk, and accepting vNM EUT does not introduce restrictions strong enough to do away with this permissiveness. Unless stronger rational restrictions than those of the orthodox vNM EUT are accepted, we are thus left with permissiveness under risk, and the rational permissibility of some non-neutral pure attitudes to risk. Before returning to the topic of CPT debiasing, note that this permissiveness of instrumental rationality under risk now creates the possibility of one problematic type of means paternalism already characterised in more general terms above. Suppose that we impose on the Cookie Monster the choices that Cookie Aficionado would have made. Despite his preference for the 47 sure cookies, we choose the 50/50 lottery on his behalf instead. Given that this is a permissible way of pursuing his cookie-eating ends (after all, we permitted this for Aficionado, who has the same ends), we are taking seriously and deferring to his subjective valuation of outcomes in this act of means paternalism. However, we are not deferring to his instrumental preferences over *how* he would like to pursue his cookie eating goals. In this case, given his own way of pursuing his ends is perfectly fine, we have no claim to pursuing his goals any more effectively than he would have himself by overriding him. And then general anti-paternalist considerations speak against overriding his instrumental preferences. This is means paternalism, but a problematic kind of means paternalism.

Returning to CPT debiasing, my argument now is that CPT debiasing is generally problematic in the same way. CPT debiasing involves identifying, for each agent, what lotteries would be favoured by an expected utility calculation using the utility function identified in a CPT model of her preferences. If this utility function is a cardinal measure of her subjective valuations of outcomes, as we are granting now (pace our arguments in the last section), then what this procedure does is impose risk neutrality with regard to subjective value on the agent. It removes any pure risk aversion or pure risk seeking inclinations the agent might have had. Given the actual permissiveness of instrumental rationality, the procedure thus imposes not *the* uniquely rational way of pursuing her ends on the agent, but rather just one of the permissible ones, the risk neutral one. And it imposes not *the* unique way of abiding by EUT in a way that respects the agent's subjective valuations, but rather just one such way. For instance, we could equally implement EUT in a way that respects her ends by letting the utility in the EUT model be a concave transformation of the CPT utility function. Under the assumption that the CPT utility function is a cardinal measure of subjective value, this would implement pure risk aversion. Despite the presence, then, of various rationally permissible options for how to pursue the agent's ends, CPT debiasing imposes just one, the one that is risk neutral with regard to subjective value. And it does so regardless of what the agent's own preferences are regarding how she would prefer to pursue her ends, expressed in her actual preferences over lotteries.

Now a defender of CPT debiasing might point to one crucial difference to the Cookie Monster case just discussed, which is that the Cookie Monster is, by hypothesis, an expected utility maximiser. The agents that CPT debiasing would be applied to are not and are thus, we have granted, in violation of the normatively correct theory of rational choice. Their CPT preferences might express their preferences regarding how to pursue their goals, but they express faulty instrumental judgements. The probability-weighting and loss aversion captured by CPT are often taken as identifying the 'mistakes' or 'biases' of agents who violate EUT, and the thought is that they justify correction.

If we accept the normative correctness of EUT, then the need for probability-weighting and loss aversion to feature in descriptively adequate models of an agent's choice behaviour is indeed indicative of a mistake — the violation of the standard axioms of EUT. But CPT debiasing seems to presuppose that they capture mistakes in a stronger sense, namely that the entire difference that probability-weighting and loss aversion make in the CPT model captures a mistake by the agent. In effect, if we also grant the CPT debiasers that CPT utility is a cardinal measure of subjective value, this presupposes that agents were trying but failing to pursue their ends in a risk neutral manner. It is hard to see where this conviction should come from. As I have argued, EUT does not require such risk neutrality. And given the actual heterogeneity in risk preferences we find using CPT and RDU representations,[19] it seems very unlikely that all agents would really be risk neutral in the pursuit of their subjective values were it not for some reasoning mistake.[20] Rather, at least some of the risk attitudes captured by probability-weighting[21] and loss aversion in a CPT model would likely remain as a pure attitude to risk were agents themselves to correct their preferences in accordance with EUT. But most importantly, we simply don't know without asking agents what their preferred way of abiding by EUT would be, out of the various permissible options, and asking them arguably obviates the need for paternalist intervention. All the means paternalist has to work with in terms of what the agent's preferred ways of pursuing her ends are are her actual CPT preferences.

Now even if risk neutrality is only one of the permissible ways of pursuing one's ends,

[19] See, for instance,l'Haridon and Vieider (2019).

[20] Also see Infante et al. (2016), who argue that there is no identifiable error in violations of EUT such as the Allais problem.

[21] Probability-weighting is sometimes treated as a merely cognitive error, whereby agents falsely represent probabilities to themselves. See, for instance, Harrison and Ross (2017). But, as mentioned above it could also represent an agent's giving more weight to comparably worse outcomes when making decisions, which can be interpreted as a source of pure risk aversion, as argued by Buchak (2013). At the very least, the interpretation of probability-weighting is essentially contestable. And my point here is that it is unlikely that it captures only mistake, because that would presuppose all agents are attempting to pursue their ends in a risk neutral way.

given we grant that the CPT preferences are irrational, we might think it is unproblematic for the policy-maker to impose one, albeit arbitrary one of the rational ways for the agent to pursue her ends. But there is a way to honour the normative authority of EUT while being much more deferential to the CPT agent's instrumental preferences over lotteries. And that is to identify the EUT model that has the closest fit with the agent's preferences over lotteries, and thus to minimise deviations from her judgements and choices.[22] The results of this procedure can be radically different from the results of CPT debiasing, as also noted by Harrison and Ross (2017), footnote 12. Suppose an agent values money linearly, and, pace the worries of the last section, the basic utility function identified by CPT reflects this linearity. CPT debiasing would then impose an EUT model with linear utility in money, and thus risk neutrality with regard to money on such an agent. However, the agent's original preferences over lotteries might have exhibited strong risk aversion with regard to money, which in the CPT model would be captured in terms of probability-weighting and loss aversion. If so, then the closest-fitting EUT model would feature a concave utility function, not a linear one.

The two procedures can thus have different results, and imposing the closest fitting EUT model on the agent strikes me as the clearly superior option on anti-paternalist grounds. Overriding an agent's preferences over means to her ends may be justifiable for those with anti-paternalist sentiments if and insofar as the policy-maker is confident she can identify a better way for the agent to pursue her ends. But whenever CPT debiasing deviates from the closest fitting EUT model, it recommends deviating further from the agent's preferences than needs be in order to secure instrumentally rational pursuit of the agent's ends. And that should be no longer acceptable for those with anti-paternalist sentiments.

6 Conclusion

Economics has traditionally been opposed to paternalism. However, the findings of behavioural economics have made popular one kind of paternalism that appears to be more innocuous: The kind of paternalism that respects an agent's ends, or her non-instrumental subjective valuations, and merely helps her pursue them effectively. CPT debiasing initially seems like a promising way to inform means paternalist policies addressed at agents who violate EUT: It allows us to identify a utility function for those agents, which is thought of by the proponents of CPT debiasing as a cardinal measure of her ends, and which we can plug into an expected utility calculation in order to determine a rational way for her to pursue those ends.

[22] This procedure is proposed, for instance, by Harrison and Ng (2016), in footnote 13.

In this paper, I have aimed to show that CPT debiasing should be opposed on general anti-paternalist grounds, even if we grant the normative authority of EUT, the descriptive adequacy of CPT, and the idea that means paternalism is at least sometimes immune to general anti-paternalist concerns. First, this is because there are reasons to doubt that the utility function measured within a CPT framework generally provides us with a cardinal measure of an agent's ends. In fact, the contestability of the psychological interpretation of the utility function alone should stop the means paternalist in her tracks.

Second, even if the utility function identified in the CPT model is a valid cardinal measure of the agent's ends, the means paternalism exhibited in CPT debiasing is a problematic type of means paternalism that should be ruled out by the same considerations that motivate economists' opposition to ordinary paternalism. This is because CPT debiasing imposes risk neutral pursuit of their ends on agents. EUT does not imply such risk neutrality, nor is such risk neutrality a plausible requirement of instrumental rationality. Plausibly, risk neutral pursuit of one's ends is just one of the permissible ways of pursuing one's ends.

In such contexts where instrumental rationality is permissive, those with anti-paternalist leanings should, as much as possible, defer to the target agent's preferences regarding how to pursue her ends. And in that case, adjustments to CPT preferences, even though we grant they are irrational, should be minimised, that is, the closest fitting EUT model should be found. CPT debiasing can involve more severe deviations from the agent's original preferences, and thus overrides an agent's instrumental preferences more than needs be in order to enforce EUT. The initial appeal of CPT debiasing seems to be based on the mistaken assumption that EUT implies risk neutral pursuit of one's ends, and that thus the entire difference that probability-weighting and loss aversion make in a CPT model captures an error on the part of the agent.

References

Mohammed Abdellaoui, Carolina Barrios, and Peter Wakker. Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory. *Journal of Econometrics*, 138(1):356–378, 2007.

Mohammed Abdellaoui, Han Bleichrodt, and Olivier l'Haridon. A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty*, 36, 2008.

Matthew D. Adler. *Measuring Social Welfare: An Introduction.* Oxford University Press, 2019.

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.

Han Bleichrodt. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics*, 11:447–456, 2002.

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.

Han Bleichrodt, Jose-Maria Abellan-Perpinan, Jose-Luis Pinto-Prades, and Ildefonso Mendez-Martinez. Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. *Management Science*, 53(3), 2007.

Adam Booij and Gijs van de Kuilen. A parameter-free analysis of the utility of money for the general population under prospect theory. *Journal of Economic Psychology*, 30(4): 651–666, 2007.

Lara Buchak. *Risk and Rationality.* Oxford University Press, 2013.

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.

Alain Chateauneuf and Peter Wakker. An axiomatization of cumulative prospect theory for decision under risk. *Journal of Risk and Uncertainty*, 18(2):137–145, 1999.

Gerald Dworkin. Paternalism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* URL = ¡https://plato.stanford.edu/archives/fall2019/entries/paternalism/¿, fall 2019 edition, 2019.

James S. Dyer and Rakesh K. Sarin. Relative risk aversion. *Management Science*, 28(8): 875–886, 1982.

Joel Feinberg. *Harm to Self.* Oxford University Press, 1986.

Daniel Groll. Paternalism, respect, and the will. *Ethics*, 122:692–720, 2012.

Till Gruene-Yanoff. Old wine in new casks: Libertarian paternalism still violates liberal principles. *Social Choice and Welfare*, 2012.

Till Gruene-Yanoff. Boosts vs. nudges from a welfarist perspective. *Revue d'economie politique*, 128(2):209–224, 2018.

Glenn W. Harrison and Jia Min Ng. Evaluating the expected welfare gain from insurance. *The Journal of Risk and Insurance*, 83(1):91–210, 2016.

Glenn W. Harrison and Don Ross. The empirical aadequacy of cumulative prospect theory and its implications for normative assessment. *Journal of Economic Methodology*, 24 (2):150–165, 2017.

Glenn W. Harrison and J. Todd Swarthout. Cumulative prospect theory in the laboratory: A reconsideration. In Glenn W. Harrison and Don Ross, editors, *Prospect Theory as a Model of Risky Choice: Descriptive and Normative Assessments*. Emerald, forthcoming.

Daniel Hausman. Behavioural economics and paternalism. *Economics and Philosophy*, 34 (1):53–66, 2018.

Daniel Hausman and Michael McPherson. Preference satisfaction and welfare economics. *Economics and Philosophy*, 25:1–25, 2009.

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.

Daniel Kahneman and Amos Tversky. Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Veronika Koebberling and Peter Wakker. An index of loss aversion. *Journal of Economic Theory*, 122:119–131, 2005.

Roman Krzysztofowicz and John Koch. Estimation of cardinal utility based on a nonlinear theory. *Annals of Operations Research*, 19:181–204, 1989.

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

Olivier l'Haridon and Ferdinand M. Vieider. All over the map: A worldwide comparison of risk preferences. *Quantitative Economics*, 10(1):185–215, 2019.

Chen Li, Zhihua Li, and Peter Wakker. If nudge cannot be applied: A litmus test of the readers' stance on paternalism. *Theory and Decision*, 76(3):297–315, 2014.

Lola Lopes. Decision making in the short run. *Journal of Experimental Psychology: Human Perception and Performance*, 9:377–385, 1981.

Lola Lopes. When time is of the essence: averaging, aspiration, and the short run. *Journal of Experimental Psychology*, 65(3):179–189, 1996.

Ben McQuillin and Robert Sugden. Reconciling normative and behavioural economics: The problems to be solved. *Social Choice and Welfare*, 38(4):553–567, 2012.

John Stuart Mill. *On Liberty.* Cambridge University Press, 2001, 1859.

Adam Oliver. The internal consistency of the standard gamble: tests after adjusting for prospect theory. *Health Economics*, 22:659–674, 2003.

Adam Oliver. Testing the internal consistency of the lottery equivalents method using health outcomes. *Health Economics*, 14:149–159, 2005.

Jonathan Parry. Defensive harm, consent, and intervention. *Philosophy and Public Affairs*, 45(4):356–396, 2017.

Richard Pettigrew. Risk, rationality, and expected utility theory. *Canadian Journal of Philosophy*, 45(5-6):798–826, 2015.

Jose-Luis Pinto-Prades and Jose-Maria Abellan-Perpinan. When normative and descriptive diverge: how to bridge the difference. *Social Choice and Welfare*, 38:569–584, 2012.

John Quiggin. A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343, 1982.

Jonathan Quong. *Liberalism without Perfection.* Oxford University Press, 2011.

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

Seanna Shiffrin. Paternalism, unconscionability doctrine, and accommodation. *Philosophy and Public Affairs*, 29:205–250, 2000.

P. F. M. Stalmeier and T. G. G. Bezembinder. The discrepancy between risky and riskless utilities: A matter of framing? *Medical Decision Making*, 19(435-447), 1999.

H. Orri Stefansson and Richard Bradley. How valuable are chances? *Philosophy of Science*, 82(4):602–625, 2015.

H. Orri Stefansson and Richard Bradley. What is risk aversion? *British Journal for the Philosophy of Science*, 70(1):77–102, 2019.

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

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

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Peter Wakker and Anne Stiggelbout. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making*, 15(2):180–186, 1995.

Douglas Glen Whitman and Mario J. Rizzo. The problematic welfare standards of behavioral paternalism. *Review of Philosophy and Psychology*, 6(3):409–425, 2015.