Behavioral Welfare Economics and Consumer Sovereignty

Guilhem Lecouteux

Behavioural welfare economics and consumer sovereignty

Forthcoming in Heilman, C. & Reiss, J. (Eds), Routledge Handbook of Philosophy of Economics

Guilhem Lecouteux

Affiliation: Université Côte d'Azur, CNRS, Gredeg, France.

Postal address: CNRS – GREDEG, Campus Azur; 250 rue Albert Einstein, CS 10269, 06905 Sophia Antipolis Cedex, France.

Telephone number: +33 (0)4 93 95 43 74

Email: guilhem.lecouteux@univ-cotedazur.fr

ORCID: 0000-0001-6602-7247

Behavioural welfare economics and consumer sovereignty

Abstract (100 words): the aim of this chapter is to critically assess the argument advanced in behavioural welfare economics that preference inconsistency and violations of rational choice theory are the result of errors, and offer a direct justification for paternalistic regulations. I argue that (i) this position relies on a psychologically and philosophically problematic account of agency, (ii) the normative argument in favour of coherence is considerably weaker than usually considered, and (iii) BWE fails to justify why agents ought to be coherent by neoclassical standards. I conclude by discussing how BWE could still justify paternalistic regulations by endorsing a more institutionalist perspective.

Keywords: behavioural welfare economics; preference inconsistency; consumer sovereignty; paternalism.

Subject classification codes: B40, D01, D60, D91

Word count: 5300 (without footnotes and references)

the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant. He cannot rightfully be compelled to do or forbear because it will better for him to do so, because it will make him happier, because, in the opinion of others, to do so would be wise, or even right. […] The only part of the conduct of any one, for which he is amenable to society, is that which concerns others. In the part which merely concerns himself, his independence is, of right, absolute. Over himself, over his own body and mind, the individual is sovereign.

It is, perhaps, hardly necessary to say that this doctrine is meant to apply only to human beings in the maturity of their faculties. […] Those who are still in a state to require being taken care of by others, must be protected against their own actions as well as against external injury. (John Stuart Mill, 2003 [1859], pp. 80-81)

1. Behavioural economics' challenge to consumer sovereignty

Consumer sovereignty constitutes a central principle in the mainstream tradition of welfare economics.[1] When assessing what is good for society, the theorist[2] [she] takes individual 'welfare' as an input, i.e. her assessment of what makes an individual [he] better off. While finding an appropriate way to aggregate measures of individual welfare into a social welfare function has been – and still is – a core challenge of theoretical welfare economics, the traditionally accepted way to measure individual welfare was to use the satisfaction of individual preferences. If an individual prefers an alternative x to y (and therefore, would choose x if asked to choose between x and y), then this individual's welfare is deemed to be higher when he obtains x.[3] Taking preference satisfaction as the normative criterion means that theorists do not form any judgement about the agents' preferences, and leave them the freedom to choose as they wish. Individual choices are taken as indicators of the agents' economic welfare, which can be used as inputs in normative analysis. This led to a consensus that normative economics is about *social interactions* (e.g. whether markets constitute a good

[1] Hutt (1940, p.66) defines the principle of consumer sovereignty as 'the controlling power exercised by free individuals, in choosing between ends, over the custodians of the community's resources, when the resources by which those ends can be served as scarce'. See Desmarais-Tremblay (forthcoming) for a historical discussion.

[2] I will use the generic term 'theorist' to refer to the actual economist, philosopher, outside observer, who intends to model a choice problem and to derive a normative judgement about it. I will occasionally use the pronoun 'we' to refer to theorists in general (I imagine that most readers could indeed find themselves in this position).

[3] I will not discuss in this chapter the relationship between welfare and preference satisfaction, and in particular whether 'welfare' should be interpreted substantively – preferences being either constitutive or providing evidence about welfare – or formally – in which case welfare is defined by the satisfaction of preferences (see Lecouteux (2021) on these questions in the context of BWE).

mechanism to allocate resources) and should consider the individuals' preferences as their protected sphere of liberty.

The development of behavioural economics however challenged this consensus. The accumulation of experimental findings that human subjects put in lab conditions are prone to preference reversals and inconsistencies indeed led a growing number of economists to question the relationship between individual choice and welfare. The standard narrative among behavioural economists is that real individuals choose 'poorly' (e.g. Sunstein 2020, p.39), and therefore that leaving them make their own choices could be harmful for themselves. In the words of Camerer *et al* (2003), who – simultaneously with Sunstein and Thaler (2003) proposal for 'libertarian paternalism' – explicitly argue that behavioural economics gives a direct justification for 'asymmetric paternalism':

> In a sense, behavioral economics extends the paternalistically protected category of 'idiots' to include most people, at predictable times. The challenge is figuring out what sorts of 'idiotic' behaviors are likely to arise routinely and how to prevent them, while imposing minimal restrictions on those who behave rationally (Camerer *et al* 2003, p.1218)

Camerer *et al* draw a distinction between two types of individuals, the 'idiots' and 'those who behave rationally'.[4] While rational agents should be let free to choose as they prefer, most real individuals should be protected against their own actions – echoing Mill's harm principle, it is as if behavioural economics revealed that most individuals are akin to children, since their behaviour in the lab indicate that they are not 'in the maturity of their faculties'. Our higher expertise, as theorists, then legitimise calls for paternalistic regulations, in the agents' own interests. The aim of *behavioural welfare economics* [BWE] consists then in looking for strategies to recover a normatively satisfactory notion of 'economic welfare' from the possibly incoherent choices of the agents. In this paper, I will focus on the contributions to BWE that emphasise the need for paternalistic regulation in the own interests of agents, and let aside the very few contributions that do not interpret deviations from rational choice as mistakes, such

[4] The distinction is common in this literature, with for instance the opposition between 'Humans' and 'Econs' for Sunstein and Thaler, or between 'Mr Spock' and 'Homer Simpson' – such dual-selves models are routinely used to contrast the 'rational' and 'psychological' parts of human agency.

as Harrison and Ross (2018) 'quantitative intentional stance' toward BWE (see Harrison (2019) and Lecouteux (2021) for overviews of the literature).

Contrary to the principle of consumer sovereignty – which treats consumer preferences as given and not subject to the theorist's scrutiny and judgement – most of the literature on BWE considers that incoherent preferences are the symptom of a deficient psychology, and therefore that they should not be integrated in welfare analysis. While a significant emphasis is put in this literature on the need to respect the true preferences of the agent – Sunstein (2014) calls it a 'means paternalism', whose aim is to help the individuals to achieve their own ends, *as judged by themselves* – behavioural welfare economists are endowed with the duty of helping non-rational agents to obtain what they 'truly' want. The preferences of an individual are worth respecting *only if he is 'rational' in the neoclassical sense*.[5] Far from a principle of 'consumer sovereignty', BWE implicitly advances a principle of 'sovereignty of the neoclassical consumer'.

The aim of this chapter is to question the widespread position among behavioural economists that incoherent preferences *do* pose a normative problem, and that preference inconsistency gives a straightforward justification for paternalistic regulations. I start by presenting in more detail how BWE analyses such incoherent preferences, and highlight the problematic account of agency it presupposes, based on the model of the *inner rational agent* (section 2). I then question the claim that revealing incoherent preferences is normatively problematic (section 3), and argue that – if we accept that people ought to reveal coherent preferences – it is far from clear why people ought to be coherent by neoclassical standards (section 4). I conclude by stressing that BWE critique of consumer sovereignty is probably misplaced, and that shifting the analysis from cognitive biases to the general processes of preference formation could offer a much more forceful argument in favour of regulation – which would however be of a very different nature than the nudge agenda (section 5).

[5] The 'proper' definition of rational, and what it means to be 'rational in the neoclassical sense', will be discussed below.

2. Behavioural welfare economics and the inner rational agent

2.1 Interpreting deviations from rational choice

Neoclassical welfare economics assumes that agents are 'rational', in the sense that (i) their preferences are complete and integrated,[6] and (ii) they act in an instrumentally rational way to satisfy those preferences. Those two assumptions can either be interpreted literally, or as a formal representation of the agent's behaviour. According to the former interpretation, the agent has stable 'tastes' and 'objectives', and has the cognitive abilities to make the best choice in any given choice situation. According to the latter however, the agent behaves *as if* (i) and (ii) were true, and her preferences are directly defined from her actual choices – see Guala (2017) on the distinction between mentalistic and behaviouristic interpretations of preferences.

Behavioural economics' challenge to welfare economics is that human subjects – when put in lab conditions – often behave very differently from the prediction of rational choice theory (see e.g. Kahneman 2011 and Camerer 2011 for references). This suggests that at least (i) or (ii) must be rejected as an empirical statement. If we reject (i), we cannot represent the agent's preferences by a utility function anymore, leaving us with no obvious way to measure – and more fundamentally, define – the subjective welfare of the agent. This raises the daunting philosophical question of which normative criterion to be used in the absence of a welfare metric. If we reject (ii) while maintaining (i) however, the agent has well-ordered preferences which can be represented by a utility function. He however fails to maximise it, because of a lack of instrumental rationality: the agent's subjective welfare is well defined (whose maximisation could offer an appealing normative criterion), although his choices now only give an *indirect* source of information (e.g. Közsegi & Rabin 2007).

In the early days of behavioural economics, Simon (1955) endorsed the first interpretation and rejected the notion of individual utility functions, while still emphasising the existence of a form of rationality – which was however procedural rather than substantial. The new behavioural economics that emerged with the 'Heuristics and biases' program of Kahneman and Tversky (see Sent 2004) – and which later became mainstream and shaped the development of BWE – predominantly endorsed the second interpretation of behavioural findings (i.e. keep (i) while rejecting (ii)). Most of the contributions to the literature in BWE treat deviations from rational choice as *errors* – caused by the defective psychology of the

[6] By *integrated* (see Sugden 2018), I mean non-stochastic, context-independent, and internally consistent – 'consistency' being defined by axioms such as transitivity or the sure-thing principle. Completeness and integration typically imply the formal conditions that allows a utility representation of preferences.

individual – and aims to reconstruct the underlying 'true' preferences of the agent, that he would have revealed, if freed from reasoning imperfections and biases.

2.2 The inner rational agent

Together with Gerardo Infante and Robert Sugden, we have argued that this literature treats human agency as if a person was made up of a neoclassically rational agent – an *inner rational agent* – 'trapped' within an error-prone psychological shell, which distorts how the inner agent interacts with the real world (Infante *et al*, 2016a,b). Our critique is not that behavioural economists think that this is a *realistic* description of the person, but rather that they consider that an actual person, if freed from reasoning imperfections, would reveal neoclassical preferences. It is assumed that the person has some latent capacity to generate complete and integrated preferences, though his many psychological biases are likely to *interfere* with this latent capacity. There is however no psychological explanation for such latent preferences (Sugden 2015), and no clear theoretical foundation for assuming that the overall integrated preferences of an agent with context-dependent preferences would be neoclassical (Krause & Steedman 1986, Lecouteux & Mitrouchev 2021). Indeed, while behavioural economists found countless theories in cognitive psychology to explain how people *deviate* from norms of rational choice, rational choice itself remains unexplained. Consider for instance time preferences: it is commonly considered in BWE that time-inconsistency is normatively problematic, and that it is the result of a deviation (by e.g. a present bias) from the 'correct' way of discounting future outcomes – exponential discounting. I will however argue below that the various explanations we could endorse to justify why discounting one's future utility is not irrational would imply behaviours compatible with hyperbolic rather than exponential discounting. If we accept that people can deviate from time neutrality (which I suggest would be the only acceptable preferences if time-inconsistent preferences are rejected), then we must *postulate* that exponential discounting is the right discounting model.

Another issue with postulating the existence of true preferences is its contradiction with the 'as judged by themselves' clause. Indeed, any deviation from the prediction of rational choice theory is explained as a violation of assumption (ii) (that agents are instrumentally rational), and not of assumption (i) (that we have complete and integrated preferences). It is not clear however why we could not imagine that persons are instrumentally rational, even though their preferences are non-standard and generate apparently incoherent choices. If one's

preferences are a matter of personal tastes, then there is little reason to expect those subjective tastes to conform to the neoclassical axioms. Take loss aversion as an illustration (see Harrison and Ross 2017 for a similar argument). Two mechanisms could explain the typical pattern of risk preferences associated with loss aversion. First, the individual can genuinely experience a higher cognitive cost when facing losses – 'utility loss aversion', captured by the parameter λ in Kahneman and Tversky (1992). Second, he can exhibit different probability weighting functions in the gain and loss domains – 'probabilistic loss aversion' (Schmidt and Zank 2008).[7] Regarding utility loss aversion, there is no straightforward reason to maintain that we can ignore the aspects of the agent's psychology that generates a relative sensitivity of losses versus gains. Regarding probabilistic loss aversion, interpreting the weighting function as a form of perceptual error and a case of wrong belief about probabilities could justify regulations designed to limit the agent's ignorance (which would be acceptable even by liberal standards). However, if we understand the weighting function as a matter of *sensitivity* to changes in probability while knowing the right probability (Fox *et al*, 2015, pp.54-55) then it is far from clear that we can ignore the aspects of the agent's psychology leading to such weighting.

To summarise thus far, BWE interprets incoherent preferences as the deviations from underlying coherent preferences. The existence of such true preferences (either actual or counterfactual) however lacks any psychological explanation, and nothing guarantees that the preferences that are supposed to represent what the individual prefers, by his own light, ought to conform to the traditional axioms of rational choice theory. I now argue that the normative argument against preference inconsistency is considerably weaker than usually recognised.

3. What is the problem (if any) with incoherent preferences?

The common justification of behavioural paternalism is that incoherent preferences are likely to generate preference reversals and may be a source of later regret. The individual would have been better off if a benevolent planner helped him earlier to make the right decisions – and given our higher expertise, we theorists are legitimised to identify the cases in which the individuals are likely to make mistakes. I see four objections to this line of argument.

[7] For didactic purposes, it is usually simpler to present loss aversion in terms of utility loss aversion – introducing probability-loss aversion indeed requires first explicating the notions of gain and loss-ranks. It should however be noted that experimental tests of cumulative prospect theory suggest that the adequate explanation is in terms of probability weighting. Harrison and Swarthout (2016) indeed review the experimental tests of cumulative prospect theory and find that in the few properly calibrated experiments, λ tends to be close to 1.

First, in a situation of preference reversal, nothing guarantees that the regrets expressed by my later self are the symptom of a mistake by my earlier self, and that improving the situation of my earlier self, *as judged by the later self*, would have been welfare-enhancing for the earlier self. The possibility that my preferences or identity may change over time, but also that we cannot *a priori* rely on the sole judgement of a later self and ignore the judgement of an earlier one, implies that regrets cannot systematically offer a justification for paternalistic interventions.[8]

Second, despite casual claims that incoherent preferences expose the individual to money-pumps, i.e. to an exploitation by malevolent third parties, the empirical evidence that incoherent preferences lead to welfare losses – or that individuals would not be able to adjust their behaviour over time to avoid such losses – is seriously lacking (see the systematic review of the literature by Arkes *et al* (2015)). Cubitt and Sugden (2001) also suggest that money-pump arguments are theoretically flawed, since a precise definition of a money-pump highlights that an invulnerability to money-pumps does *not* require exhibiting coherent preferences by neoclassical standards. We could therefore have individuals invulnerable to money-pumps, while exhibiting non-standard preferences.

Third, the empirical evidence that individuals make incoherent choices are based on choices realised in the controlled environment of a lab experiment. Now, in most real-life settings, the uncertainty faced by decision makers is much more radical (people are rarely asked in their daily life to choose between different prospects with known probabilities). Being 'irrational' in the lab therefore gives little evidence that the agent is not rational outside the lab. Vernon Smith formulated this concern as follows (in an unpublished letter to Harsanyi, 1989):[9]

> Another issue that has long bothered me in interpreting "violations" of vNM utility is the following: decision makers are accustomed to making decision in environments in which there is uncertainty about how many states there are, an uncertainty as to the description of every possible state. We bring subjects into the laboratory where we put them in environments in which we can guarantee what the alternative states are, and that

[8] I have developed this argument in more detail about retirement savings and time-inconsistent preferences in Lecouteux (2015).
[9] I am very grateful to Dorian Jullien for sending me a scan of this letter during his stay at Duke University.

> the set is exhaustive. To what extent do people make "mistakes" in the latter environment because there [sic] intuition is programmed for the former? … For example, people tend to overweight the likelihood (sample) relative to their priors in well-defined Bayesian "learning" experiments. Well, this makes sense intuitively if the sample is a major source of learning about how rich is the set of states! Its like you had just drawn a green ball from an urn thought to contain only black and red balls.

If we want to express normative judgements about how people behave in real life, we need a realistic model of the individual's cognition, because the 'computations of a model of cognition need to be tractable in the real world in which people live, not only in the small world of an experiment with only a few cues' (Gigerenzer *et al* 2008, p.236). Optimisation models and Bayesian updating are very relevant in the small world of a controlled lab experiment – and experimental subjects frequently deviate from their theoretical predictions – though they are inadequate in fundamentally uncertain large worlds,[10] for which simple heuristics might be normatively more relevant (Gigerenzer and Sturm 2012, pp.262-264).

Last but not least, we should also question how 'obvious' is the general normative argument against inconsistency. A likely bias is that theorists – the actual persons thinking about those questions – also tend to excessively value the importance of consistency. Nozick (1981, p.407) suggests for instance that 'philosophers are people with very strong motivations to avoid inconsistency', mostly because they are a self-selected group of people – by their origins and training – who value consistency highly. This is probably even truer of behavioural economists, who – because of their training in economics and their use of mathematical models and Bayesian techniques – have been taught for years that inconsistency was highly problematic, and that experimental economics discovered 'anomalies' and 'deviations' from the norm of rational choice. There is however no straightforward ethical argument (apart from an explicit endorsement of some form of epistocracy) in favour of entrusting a group of people with PhDs in economics or philosophy, who are sociologically very far from being representative of the general population, with the task of defining what kind of life people in general could normatively desire. As Sugden (2006, p.50) puts it:

[10] I discuss in detail this distinction in terms of small and large worlds (in Savage's sense) in Lecouteux (2021).

> When political philosophy is written from the stance of the moral observer, the reality of these risks is too easily overlooked. In proposing his own conception of what is valuable, an author has to provide a reasoned defence of his position. In doing this, it is easy to slip into assuming that anyone who understands these reasons will find them convincing. Without noticing, we can make the transition from the belief that we are right to the belief that we will come out on the winning side of a reasoned discussion about what is right. So, we are inclined to think, we have nothing to fear from allowing evaluative issues to be resolved in a properly conducted democratic process. Indeed it is surprisingly easy to go further, and to imagine that the process gas already been carried out, and everyone *has* agreed with us (Sugden 2006, 50, italics in original)

Even though our own training and background as theorists – and then our own convictions as citizens about what constitutes a good life – might let us consider that consistency is of utmost importance, respecting the 'as judged by themselves' clause should also mean respecting the right of other people to act irrationally.

4. What are coherent preferences?

Suppose now, for the sake of the argument, that incoherent preferences do pose a normative problem. The question that directly follows is which criterion of 'coherence' should be used to define what is the 'right' behaviour. If we look at the literature in behavioural economics, the typical deviations documented by lab experiments are deviations with respect to (i) social preferences, (ii) time preferences, and (iii) risk preferences. Welfare-relevant preferences are supposed to be self-interested, time preferences consistent with *exponential discounting*, and risk preferences consistent with *subjective expected utility theory*. Being coherent 'by neoclassical standards' means respecting the conditions listed here. My aim in this section is to highlight that different norms of consistency may accommodate deviations from neoclassical standards. This means that we could very well be coherent with respect to a certain standard, while still exhibiting what a neoclassical theorist would consider as preference inconsistency.

4.1 Social preferences

Regarding social preferences, I will not discuss here the problems that may arise from double counting utilities in welfare measures, and whether we should exclusively consider our own counterfactual self-interested welfare, purified from other-regarding concerns (as suggested by e.g. Hausman 2012). We can however keep the focus on questions of choice consistency, by noting that neoclassical consistency requires agents to act systematically on the same preferences. This means that if I am apparently prosocial with an anonymous partner in a lab experiment (e.g. I give a significant share of my endowment in a dictator game), then it is also assumed that I will continue to be prosocial in the rest of the experiment, and possibly outside the lab too.

It would however be perfectly sensible to imagine that I could be prosocial in certain environments, with certain partners, and under specific circumstances, while selfish in many other cases. If we model sociality with an intention-based rather than outcome-based model, subjects could apparently 'switch' between selfish and prosocial preferences depending on the circumstances. This would be considered as a case of preference inconsistency by neoclassical standards, but not necessarily within the framework of e.g. psychological games (Geanakoplos *et al* 1989, Battigali & Dufwenberg 2009 – see Battigalli *et al* 2019 for an overview) or team reasoning (Sugden 1993, Bacharach 2006 – see Lecouteux 2018 for an overview). The question that arises here is whether we can legitimately ignore prosocial *intentions* in normative analysis – if not, then we should refer to a standard of consistency that accommodates such intentions, which is not true of BWE.

4.2 Time preferences

Consider now time preferences. Exponential discounting means that the individual uses a constant discount rate over time, which guarantees that his choices are time-consistent. Now, an interesting question would be to know whether neoclassical rationality requires a specific value for this discount rate. When considering the agent over time, if we assume that 'all parts of one's future are also parts of oneself; that there is a single, enduring, irreducible entity to whom all future utility can be ascribed' (Frederick 2003, p.90) – which seems to be the case in neoclassical analysis of time preferences, with the assignment of an *undated* utility function to the agent – then there is no decisive argument for discounting future utilities. Temporal

neutrality, with an equal weighting of all time periods, is for instance explicitly endorsed by O'Donoghue and Rabin (1999).

I have argued elsewhere that, if we accept the argument that time-inconsistent choices reveal a 'mistake', then people *ought to be time neutral* (Lecouteux 2015). This is a rather strong normative claim, and there are many reasons why people could legitimately discount future outcomes. It is worthy to note however that, if we consider the various reasons that could justify – from a normative perspective – discounting the future, it seems that the agent should *not* use a constant discount rate.

A first reason for discounting one's future utilities is the uncertainty of the future, and that agents have a noisy estimation of their future utilities (because of a limited ability to foresee future experiences). Even if we consider an unbiased noise, Gabaix and Laibson (2017) show that the resulting behaviour is consistent with hyperbolic rather than exponential discounting. Another related motive that could justify discounting the future is the possibility of dying (or at least of not being able to collect outcomes in the future). However, this probability will not be constant over time, which will justify the use of a non-constant discount rate. A third motive for discounting one's future utilities would be that my preferences and identity are likely to evolve over time, as in Parfit's (1984) complex view of identity. This however also generates a behaviour consistent with hyperbolic rather than exponential discounting (Lecouteux 2015). Lastly, we can consider the opportunity cost of time in terms of financial savings. We should however again use a constant discount rate only if interest rates were themselves constant over time (which is obviously not the case).

4.3 Risk preferences

Consider finally risk preferences. When evaluating a prospect with known probabilities, an agent whose choices are consistent with expected utility maximisation behaves as if he had a linear valuation of probabilities, and a strictly increasing – not necessarily linear – valuation of outcomes. The first problem is that it would be perfectly sensible to define conditions for coherent preferences with a non-linear valuation of probabilities, such as in rank-dependent utility (see Wakker 2010 for a detailed discussion and an axiomatisation), or the closely related risk-weighted expected utility of Buchak (2013). Those models can indeed be axiomatized while only slightly relaxing the conditions under which expected utility holds – the critical condition being unrestricted tradeoff consistency, and there are good normative arguments in

favour of weaker versions such as rank-tradeoff consistency or comonotonic tradeoff consistency.

A second problem is the distinctive treatment of probabilities and outcomes in subjective expected utility theory. Indeed, for the same reason that increasing your income by a fixed amount has a different marginal impact whether you are a beggar or a millionaire (captured by the degree of concavity or convexity of your utility function), increasing the probability of occurrence of the best outcome by 1 percentage point will have a different marginal impact on your preferences depending on the initial level of the probability (the impact will be significant if the initial probability is 0% or 99%, though it is likely to be negligible for e.g. 40%). Descriptively, we tend to perceive both outcomes and probabilities non-linearly. Normatively however, there is no obvious reason why we ought to treat probabilities linearly, *and not outcomes*. This leads to an inconsistency in neoclassical rationality: if we allow for a non-linear perception of outcomes – and therefore allows for deviations from risk neutrality – then the agent's preferences contain a Dutch book (de Finetti 1931, see also Wakker 2010, chap.1).[11] This means that if the possibility of being exploited by a third party is the symptom of preference inconsistency, then neoclassical rationality should require being risk neutral.

Even though we accept the view that our preferences ought to be coherent, the position that we ought to be coherent by neoclassical standards is rather weak. Letting social preferences aside, it seems that a strict understanding of what it means to be coherent should imply time neutrality (any motive that could justify discounting the future would indeed also justify hyperbolic discounting) as well as risk neutrality (since it is the only risk attitude protecting us from a Dutch book). There are however other standards of consistency that seem normatively acceptable, though they will generate behaviours inconsistent with neoclassical standards. Before pointing out to people's 'deviations', BWE should first justify why we ought to be (i) self-interested, (ii) time-consistent, and (iii) expected utility maximisers.

[11] The intuition is the following. Consider two complementary prospects such that P_E pays you 1 if and only if E happens, and $P_{\bar{E}}$ pays you 1 if and only if E does not happen. If you are (say) risk averse, you will value both prospects at less than their expected value. Someone who buys from you P_E and later $P_{\bar{E}}$ would therefore spend less than 1. However, the same individual would then be able to sell you at a price of 1 the two prospects bundled together (since $P_E \cup P_{\bar{E}}$ pays 1 for sure). You would then end up with a sure loss.

5. Concluding remarks: changing BWE's lens

I have argued in this chapter that contrary to the conventional wisdom of BWE, the fact that subjects put in the lab violate the standards of (neoclassical) rational choice should *not* be interpreted as a mistake on the agent's behalf. Such evidence therefore does not offer a straightforward argument in favour of paternalistic regulations. BWE argument indeed relies on the validity of the model of the inner rational agent, it probably overestimates the normative appeal of consistency, and it fails to properly justify why the 'correct' way of behaving should be consistent with neoclassical standards.

I would like to conclude this chapter by emphasising that my argument is not against paternalism *per se*, but rather against the *justification* for paternalism advanced in BWE. While I do not consider incoherent preferences to be fundamentally problematic, BE highlights one important point that could justify regulations: the fact that people's preferences could be manipulated by third parties for their personal, commercial or political interests – unlike the benevolent nudgers and choice architects central to BWE. The normative problem does not lie in the preferences, but in the process of preference formation. This point echoes Galbraith's (1938) early critique of the consumer sovereignty principle, who similarly rejected the idea of a difference between 'rational' and 'irrational' preferences put forward by Kahn (1935). According to Chirat (2020), Galbraith rather emphasised the *endogeneity* of preferences, and that 'the formation of preferences does not lie in the inner rational individual but in a cultural scheme and social interactions' (Chirat 2020, p.267). If individual preferences are the product of a social system, it is problematic to use them as the fundamental building block of welfare analysis. BWE should probably acknowledge both the individual and social determinants of our preferences and behaviours, rather than merely blaming the defective psychology of individual agents taken in isolation. Endorsing the view that our preferences are fundamentally *shaped* by the social environment and the supply-side of the market[12] would probably justify more ambitious social policies than nudges, whose aim is to correct *ex post* anomalies in our behaviours, rather than to tackle *ex ante* the causes of such anomalies.

[12] In the context of addictive behaviours, Ross (2020) argues that becoming un-addicted is mostly a question of improving one's heuristics (and therefore, a matter of individual decisions), while addiction is the outcome of socially engineered addictive environments. An adequate policy against addictive behaviours should therefore help individuals to quit (thanks to e.g. boosts) but also address the fact that the business model of some industries is precisely to foster addiction, such as the cigarette industry or social networking sites.

References

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence?. *Decision*, *3*(1), 20.

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, *144*(1), 1-35.

Battigalli, P., Corrao, R., & Dufwenberg, M. (2019). Incorporating belief-dependent motivation in games. *Journal of Economic Behavior & Organization*, *167*, 185-218.

Buchak, L. M. (2013). *Risk and rationality*. Oxford University Press.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C., Issacharoff, S., Loewenstein, G., O'donoghue, T., & Rabin, M. (2003). Regulation for Conservatives: Behavioral Economics and the Case for" Asymmetric Paternalism". *University of Pennsylvania law review*, *151*(3), 1211-1254.

Chirat, A. (2020). A reappraisal of Galbraith's challenge to Consumer Sovereignty: preferences, welfare and the non-neutrality thesis. *The European Journal of the History of Economic Thought*, 27(2): 248-275.

Cubitt, R. P., & Sugden, R. (2001). On money pumps. *Games and Economic Behavior*, *37*(1), 121-160.

De Finetti, B. (1931). Sul significato soggettivo della probabilita. *Fundamenta mathematicae*, *17*(1), 298-329.

Desmarais-Tremblay, M. (forthcoming). WH Hutt and the conceptualization of consumers' sovereignty. Forthcoming in *Oxford Economic Papers*.

Fox, C. R., Erner, C., & Walters, D. J. (2015). Decision under risk: From the field to the laboratory and back. *The Wiley Blackwell handbook of judgment and decision making*, 43-88.

Frederick, S. (2003). Time preference and personal identity. *Time and decision*, 89-113.

Gabaix, X., & Laibson, D. (2017). *Myopia and discounting* (No. w23254). National bureau of economic research.

Galbraith, J. K. (1938). Rational and irrational consumer preference. *The Economic Journal*, *48*(190), 336-342.

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, *1*(1), 60-79.

Gigerenzer G, Hoffrage U, Goldstein DG. 2008. Fast and frugal heuristics are plausible models of cognition: reply to Dougherty, Franco-Watkins, and Thomas. *Psychol. Rev*. 115: 230–39.

Gigerenzer, G., & Sturm, T. (2012). How (far) can rationality be naturalized?. *Synthese*, *187*(1), 243-268.

Guala, F. (2017). Preferences: neither behavioural nor mental. *Economics & Philosophy*, 1-19.

Harrison, G. W. (2019). The behavioral welfare economics of insurance. *The Geneva Risk and Insurance Review*, *44*(2), 137-175.

Harrison, G., & Swarthout, T. (2016). Cumulative prospect theory in the laboratory: A reconsideration.

Harrison, G. W., & Ross, D. (2017). The empirical adequacy of cumulative prospect theory and its implications for normative assessment. *Journal of Economic Methodology*, *24*(2), 150-165.

Hausman, D. M. (2012). *Preference, value, choice, and welfare*. Cambridge University Press.

Hutt, W. H. (1940). The concept of consumers' sovereignty. *The Economic Journal*, 66-77.

Infante, G., Lecouteux, G., & Sugden, R. (2016a). 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.

Infante, G., Lecouteux, G., & Sugden, R. (2016b). 'On the Econ within': a reply to Daniel Hausman. *Journal of Economic Methodology*, *23*(1), 33-37.

Kahn, R. F. (1935). Some notes on ideal output. *The Economic Journal*, *45*(177), 1-35.

Kahneman, D. (2011).Thinking, fast and slow. New York, NY: Farrar, Straus and Giroux.

Krause, U., & Steedman, I. (1986). Goethe's Faust, Arrow's Possibility Theorem and the individual decision taker. In Elster, J (Ed) *The multiple self*, 197-231.

Kőszegi, B., & Rabin, M. (2007). Mistakes in choice-based welfare analysis.American EconomicReview, 97, 477–481.

Lecouteux, G. (2015). In search of lost nudges. *Review of Philosophy and Psychology*, *6*(3), 397-408.

Lecouteux, G. (2018). What does "we" want? team reasoning, game theory, and unselfish behaviours. *Revue d'économie politique*, *128*(3), 311-332.

Lecouteux, G. (2021). Welfare economics in large worlds: welfare and public policies in an uncertain environment, In Kincaid, H. & Ross, D. (Eds), *Elgar Modern Guide to the Philosophy of Economics.*

Lecouteux, G., & Mitrouchev, I. (2021). The 'View from Manywhere': Normative Economics with Context-Dependent Preferences. GREDEG Working paper

Mill, J. S. (2003) [1859]. *On Liberty*, reprinted in Bromwich and Kater (Eds), *Yale University Press*.

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American economic review*, *89*(1), 103-124.

Parfit, D. (1984). *Reasons and persons*. OUP Oxford.

Ross, D. (2020). Addiction is socially engineered exploitation of natural biological vulnerability. *Behavioural Brain Research*, 112598.

Schmidt, U., & Zank, H. (2008). Risk aversion in cumulative prospect theory. *Management Science*, *54*(1), 208-216.

Sent, E. M. (2004). Behavioral economics: how psychology made its (limited) way back into economics. *History of political economy*, *36*(4), 735-760.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99-118.

Smith, V. (1989). Letter to John Harsanyi, October 12, 1989. In Vernon Smith papers, Correspondence, Box number 14, 1989 June-Dec (Folder 2 of 3); at David M. Rubenstein Rare Book and Manuscript Library, Duke University.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social philosophy and policy*, *10*(1), 69-89.

Sugden, R. (2006). What we desire, what we have reason to desire, whatever we might desire: Mill and Sen on the value of opportunity. *Utilitas*, *18*(1), 33-51.

Sugden, R. (2015). Looking for a psychology for the inner rational agent. *Social Theory and Practice*, *41*(4), 579-598.

Sugden, R. (2018). *The community of advantage: A behavioural economist's defence of the market*. Oxford University Press.

Sunstein, C. R. (2014). *Why nudge?: The politics of libertarian paternalism*. Yale University Press.

Sunstein, C. R. (2020). *Behavioral Science and Public Policy* (Elements in Public Economics). Cambridge: Cambridge University Press.

Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review*, 1159-1202.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation ofuncertainty.Journal of Risk and Uncertainty, 5, 297–323.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge university press.