Welfare economics in large worlds: welfare and public policies in an uncertain environment

Guilhem Lecouteux

Welfare economics in large worlds: welfare and public policies in an uncertain environment

Forthcoming in Kincaid, H & Ross, D. (Eds), *A Modern Guide to Philosophy of Economics*, Elgar

Guilhem Lecouteux

Université Côte d'Azur, CNRS, GREDEG, France.

Postal address: 250, rue Albert Einstein, 06560 Valbonne, France. Email: guilhem.lecouteux@univ-cotedazur.fr

Abstract (124 words): the aim of this chapter is first to review the different approaches to the problem of reconciling normative economics with the empirical findings of behavioural economics. I distinguish between welfarist, behaviourist, constitutional, and procedural approaches, depending on whether we endorse or reject preference satisfaction as a valid descriptive or normative statement. I then argue that Savage's distinction between small and large worlds offers the adequate framework to conceptualise the problem of inferring a notion of welfare from possibly incoherent choices. I show that the four types of approaches offer complementary solutions to the reconciliation problem, depending on the nature of the uncertainty of the choice problem, and on the epistemic position of the theoretician relative to that of the agent we intend to model.

Keywords: reconciliation problem, behavioural welfare economics, nudge, boost, large worlds, welfare

JEL codes: A11 B40 D01 D63 D91

Word count: 9500 (without references and footnotes).

Welfare economics in large worlds: welfare and public policies in an uncertain environment

> The undeniable success and growing influence of behavioural economics presents us with an important challenge. Normative questions are central to economics. For well over half a century, the dominant approach to those questions has been rooted in the paradigm of revealed preferences, which instructs us to infer objectives and welfare … from choices. But behavioral economics teaches us that choices are not always consistent … How can we make coherent statements about welfare when the choices to which we look for guidance are inconsistent for reasons we do not fully understand?
>
> Bernheim 2016, pp.12-13

> 'Why should a rational decision-maker wish to be consistent? After all, scientists aren't consistent, on the grounds that it isn't clever to be consistently wrong. When surprised by data that shows current theories to be in error, they seek new theories that are inconsistent with the old theories. Consistency, from this point of view, is only a virtue if the possibility of being surprised can somehow be eliminated. This is the reason for distinguishing between large and small worlds. Only in the latter is consistency an unqualified virtue.'
>
> Binmore 2007, p.28

Introduction

Neoclassical welfare economics is founded on a tight conceptual connection between preferences, choices, and welfare. C*hoices* constitute the primitive of analysis (as they are directly observable by the theoretician[1]), which are encoded in a *preference relation* over the alternatives. If it respects certain formal consistency requirements, this preference relation can itself be represented by a *utility function*. An individual's *welfare* is then inferred from the satisfaction of her preferences. It is thus strictly equivalent to state:

(i) the welfare of the individual is higher with x than with y
(ii) x is preferred to y
(iii) If given the opportunity to choose between x and y, the individual would choose x

It should however be emphasised that the formulation above is often confused with a more 'psychological' version of welfare, according to which the maximisation of welfare is the *explanation* of the agent's choice. That is, instead of a logical relation in terms of equivalence

[1] I will use the generic term 'theoretician' to designate a behavioural/welfare economist who intends to model a choice problem, and infers the preferences and welfare of the agent from her theoretical model.

(i) ⇔ (ii) ⇔ (iii) it is common practice among economists to suppose that the logical relation between welfare and choice is an implication of the form (i) ⇔ (ii) ⇒ (iii). To avoid any confusion, I will refer to the former version in terms of equivalence as 'mindless' welfare economics and to the latter as 'mindful' welfare economics.[2]

The accumulation of experimental findings that human subjects' choices in the lab significantly differ from the prediction of neoclassical economics[3] led many behavioural economists to question this connection between choice and welfare. Indeed, the main narrative that progressively pervaded the discourse of behavioural economists is that individuals regularly take decisions that contradict their 'best interest' (e.g. Thaler and Sunstein 2008, p.9). More accurately, the challenge from behavioural economics is that individuals may reveal *inconsistent* preferences, from which it might be challenging to infer normative judgements about their choices. This literature in behavioural economics led to the development of a research program in *behavioural welfare economics* [henceforth, BWE], which looks for strategies to recover a normatively satisfactory notion of 'economic welfare' from the possibly inconsistent choices of the agents.

While most[4] contributions to BWE discuss how agents *deviate* from the satisfaction of their preferences, my aim in this chapter is to shift the focus from the agents' supposed cognitive deficiencies to a more systemic analysis of choices in *large worlds*. Explicitly invoking Savage's distinction between small and large worlds can indeed contribute to significantly clarify the debate about the 'correct' definition of welfare in BWE – with distinct approaches interpreting preferences as mental states (requiring an analysis of the *actual* process of decision making) or stable patterns of behaviour (for which a *Bayesian representation* of choices is sufficient). I will argue in particular that most behavioural welfare economists mix those two approaches by interpreting the Bayesian framework as an idealisation rather than representation of actual processes of decision-making.

I start by recalling some foundational notions of neoclassical and behavioural welfare economics, and emphasise a common confusion about the meaning and proper interpretation of preferences and welfare (section 1). I continue by briefly reviewing the different solutions to the problem of reconciling normative and behavioural economics, and propose a classification based on their respective positions *vis-à-vis* the interpretation of preferences for descriptive and normative purposes (section 2). I propose a reformulation of the debate in terms of choices in large worlds, and highlight how the different contributions to the reconciliation problem can be interpreted within this framework (section 3). I then argue that the choice of the relevant framework for measuring welfare and designing policies depends first on the nature of the uncertainty of the choice problem, and second on our epistemic position as theoreticians relative

[2] I borrow the two terms from Gul and Pesendorfer (2008) and Camerer (2008) respective arguments for a mindless and mindful economics, whose disagreement is founded on the very same distinction between preferences as representing or explaining choices.

[3] See e.g. Camerer (2011) and Kahneman (2011) for references.

[4] The significant exceptions to this approach are discussed in detail later in this chapter.

to that of the agent, i.e. whether we expect to know more or less about the choice problem than the agent we intend to model and advise (section 4).

1. Neoclassical welfare economics and the challenge from behavioural economics

In neoclassical economics, both positive and normative analysis attribute to economic agents consistent and context-independent preferences (Hausman 2012, p.26). Consistency is usually defined by the weak axiom of revealed preferences,[5] which entails that the choices of the agents can be represented by a complete and transitive preference relation. Context-independence means that the individual's preferences, in addition to being internally consistent, should also remain stable across different contexts.[6] Given this notion of preferences, I suggest that the core of neoclassical welfare economics can be captured by the two following statements:

> **Statement (B)** (behavioural assumption): individuals choose what they prefer
> **Statement (N)** (normative assumption): it is good that individuals get what they prefer

From (B) and (N), we can deduce the principle of *consumer sovereignty,* which is central in neoclassical welfare economics (e.g. Robinson 1979, p.92), and which could be stated as 'it is good to let individuals make their own choices'. The interpretation of the two statements (B) and (N) – and then, the normative justification of the principle of consumer sovereignty – however crucially depends on how we interpret the expression 'what they prefer'.

While mindless welfare economics is built on a *behaviouristic* interpretation of preferences (according to which preferences are a mere representation of choices), mindful welfare economics starts from a *mentalistic* interpretation of preferences, i.e. preferences as mental states that cause the agents' choice.[7] As an illustration of those two practices, consider Graaff's 'classical' (in Samuelson's own words) definition of welfare:

> The matter can be put somewhat formally by saying that a person's welfare map is defined to be identical with his preference map – which indicates how he would choose between different situations, if given the opportunity for choice. To say that his welfare would be higher in *A* than in *B* is thus no more than to say that he would choose *A* rather than *B*, if he were allowed to make the choice. (Graaff 1957, p.5)

And then his concern about the proper interpretation of the term 'utility':

> The function describing a man's hypothetical choices can be called his *utility* function. It should be emphasised that it *describes* choices, and in no way seeks to *explain* them. […] Its name is, for historical reasons, rather unfortunate. It tempts people to ask if 'utility' is

[5] Simply put, the axiom states that, if an agent chooses *x* when *y* is available, then the agent should never choose *y* from a set of alternatives that includes *x*.

[6] While the notion of context is routinely used in behavioural economics and seems rather intuitive – as e.g. the 'circumstances', 'background', or 'irrelevant features' of the choice situation – looking for a precise characterisation easily leads to circular definitions. I put aside for the moment the question of the adequate definition of 'context', which will be clarified in section 3 in terms of *small world representations*.

[7] For a detailed discussion on the behaviouristic and mentalistic interpretations of preferences, see Guala (2019).

> measurable. If we spoke instead of a 'hypothetical choice function', few would wonder if 'hypothetical choice' were measurable (Graaff 1957, p.34)

Despite this warning, many economists usually endorse an interpretation based on 'well-being',[8] and see preferences as the cause of behaviours. Bernheim (2016, p.16) for instance writes:

> [Standard welfare economics] invokes general premises that associate welfare with choices. I would articulate them as follows:
> *Premise 1: Each of us is the best judge of our own well-being*
> *Premise 2: Our judgments are governed by coherent, stable preferences*
> *Premise 3: Our preferences guide our choices: when we choose, we seek to benefit ourselves*

Under a behaviouristic interpretation, (B) is a tautology (since preferences are meant to represent choices, agents choose what they prefer *by definition of their preferences*), and (N) means that it is desirable to let people choose whatever they would choose, if given the opportunity for choice. The normative stance of mindless welfare economics is not about the agents' mental states (such as happiness or well-being), but about *freedom of choice.* Robinson summarises this interpretation as follows:[9]

> We are told nowadays that since *utility* cannot be measured it is not an operational concept, and that 'revealed preferences' should be put in its place. Observable market behaviour will show what an individual chooses. Preference is just what the individual under discussion prefers; there is no value judgment involved. Yet, as the argument goes on, it is clear that it is a Good Thing for the individual to have what he prefers. This, it may be held, is not a question of satisfaction, but freedom – we want him to have what he prefers so as to avoid having to restrain his behaviour. (Robinson, 1974 [1962], p.50)

Under a mentalistic interpretation however, (N) is an endorsement of welfarism – the principle that the goodness of a state of affairs is based on the level of an objective or subjective notion of well-being. (B) is then a genuine empirical assumption about how people choose, stating that the individual manages to choose optimally to maximise her welfare.

As noted by McQuillin and Sugden (2012, p.555), as long as economists could assume that individuals' preferences were consistent and context-independent, they 'could use a common theoretical system in which preference-satisfaction was the normative standard, while disagreeing about *why* preference-satisfaction mattered'. Behavioural economics, by challenging the assumption of coherent preferences, however forces economists to clarify their interpretation of preferences, and more fundamentally their normative positions. Under a behaviouristic interpretation, (B) remains a tautology, though (N) may now be disputed. This is for instance Sen's view on the problem of adaptive preferences, according to which endogenous preferences might

[8] Hausman (2012) for instance explicitly acknowledges using 'welfare' and 'well-being' as synonymous.
[9] This interpretation of preference satisfaction in terms of freedom of choice is also explicit in Sugden (2004), Bernheim and Rangel (2007, p.464), and McQuillin and Sugden (2012).

hamper the *effective* freedom of the agent.[10] Under a mentalistic interpretation, the fact that people's choices are not consistent does not challenge (N) as a normative stance, but rather the empirical validity of (B). This is for instance the approach favoured by Sunstein and Thaler (and by most behavioural economists), with an emphasis on the possibility of human fallibility:

> we clearly do not always equate revealed preference with welfare. That is, we emphasize the possibility that in some cases individuals make inferior choices, choices that they would change if they had complete information, unlimited cognitive abilities, and no lack of willpower (Thaler and Sunstein, 2003, p.175)

McQuillin and Sugden labelled the 'reconciliation problem' the question of how to reconcile normative economics with the findings of behavioural economics. In their initial review of the literature in 2012, they distinguished between three different solutions, based on three different interpretations of the criterion of preference satisfaction – as the maximisation of 'objective' well-being, the satisfaction of subjective welfare, or as consumer sovereignty. Given the significant volume of contributions to this debate in recent years (see also Harrison 2019, §6 for a recent overview of the literature), I suggest a different lens to interpret the different strategies endorsed by behavioural economists, namely their position towards the statements (B) and (N). The table below summarises the different approaches I will discuss:

	Keep (B)	Reject (B)
Keep (N)	Behaviourist	Welfarist
Reject (N)	Constitutional	Procedural

It is important to note that each category listed here may be constituted of very heterogenous contributions: they are however unified by their distinctive approach to the reconciliation problem. *Welfarist* approaches are in the direct continuation of mindful welfare economics, and emphasise how the agent's cognitive deficiencies prevent her from maximising her welfare. *Behaviourist* approaches, on the other hand, maintain the tradition of mindless welfare economics – considering choices as the primitive for normative analysis – while defining more precisely the scope of validity of welfare analysis. *Constitutional* approaches also keep a behaviouristic interpretation of preferences, though they shift the normative focus from preference satisfaction to the institutional environment of the choice problem. Lastly, *procedural* approaches keep the perspective of mindful welfare economics seeking to explain rather than merely representing choices, while shifting the normative focus from outcomes to the *process* of decision-making.

[10] Sugden rejects this argument, and prefers a notion of opportunity in terms of *negative* freedom (Sugden (2006, pp. 46-47)), which does not require any specification of the agent's preferences.

2. A review of the solutions to the reconciliation problem

2.1. Welfarist approaches

The first approach, which I will call the *welfarist* solution to the reconciliation problem, is in the direct continuation of mindful welfare economics. It interprets preferences as mental states, and argues that the evidence put forward by behavioural economics invalidate (B). Given that (N) remains a valid normative criterion, the challenge is to define a standard of welfare that cannot directly be inferred from the agents' choices.

A strategy would be to postulate *a priori* what counts as 'welfare', such as specific mental states – e.g. 'pleasure' in the spirit of hedonism, which is explicitly endorsed in the Kahneman *et al* (1997) 'Back to Bentham' strategy. This is the approach of the economics of happiness literature (Layard 2011, Diener and Diener 2008, Frijters *et al* 2020), though it is decisively problematic on many accounts.[11] A related strategy consists in defining an objective list of what contributes to well-being, such as friendships or the development of one's capacities – see e.g. Griffin (1986) or Nussbaum's (2011) approach to capabilities. The obvious difficulty of such *a priori* accounts is the arbitrary definition of welfare, which is difficultly reconcilable with welfare economists' general reluctance to venture into contested philosophical terrain – reluctance which motivated the tradition of defining economic welfare as an *ex post* notion inferred from individuals' choices.

Another possibility, which seems at first glance to avoid defining welfare *a priori*, is to start from choice data, and to try to *purify* the preferences of the agent so as to recover her underlying true preferences. Starting from choice data should indeed allow theoreticians to respect people's views about their welfare, 'as judged by themselves', in the often-quoted words of Thaler and Sunstein (2008). This is for instance the position advocated by Hausman (2012, 2016, 2020) according to whom even though preference satisfaction does not equate to welfare, it still offers an (imperfect) indicator of welfare. This argument however implicitly presupposes that, if the individual were free from 'reasoning imperfections' (i.e. not affected by the many biases listed by the heuristics and biases research program), she would reveal preferences that respect the standard requirements of internal consistency and context-independence. Modelling the individual as this kind of inner rational agent, impaired by the cognitive biases of an outer psychological shell, is however psychologically unfounded,[12] and philosophically problematic (Infante *et al* 2016a,b). The literature relying on the model of the inner rational agent,[13] despite its claim of respecting the

[11] See e.g. Tiberius and Plakias, 2010, pp.405-407 for a philosophical discussion, and Singh and Alexandrova 2020 for a critical comment on the political implications of the approach.

[12] It must be emphasised that the metaphor of the inner rational agent is not a claim about the *ontological commitment* of behavioural welfare economists, but a claim about their belief in the existence of some kind of mode of latent reasoning *that would generate neoclassical preferences*. In the absence of arguments supporting that a 'reasoning free from psychological imperfections' (which incidentally requires an algorithm rather than a human) would necessarily lead to complete, transitive, and context-independent preferences, we should question even the *counterfactual* possibility of such an inner rational agent.

[13] Literature that we labelled 'behavioural welfare economics' with Gerardo Infante and Robert Sugden, although I now prefer to keep the label BWE for a wider set of works – including in particular the behaviourist approach discussed below.

agent's subjective preferences, therefore imposes consistency and context-independence as *a priori* requirements for the agent's true preferences. Just as in the hedonistic approach, what counts as welfare is defined *a priori*, independent from the agent's actual choices. The ends of the individual are replaced by the ends of a counterfactual individual, her inner rational agent – and since the theoretician is the sole judge of how to adequately purify the preferences of the agent, there is a risk that the theoretician imposes *in fine* her own views about what she thinks it is rational to desire.[14]

2.2 Behaviourist approaches

The second alternative, the *behaviourist approach* to the reconciliation problem, shares significant modelling similarities with the welfarist solution, though its interpretation is much more in line with the tradition of mindless welfare economics. Individual choices are considered as the valid primitive for normative analysis, rather than an *a priori* notion of welfare. The strategy of the behaviourist solution is to maintain *both* statements (B) and (N), and to use the standard representation tools of Bayesian analysis to recover the agents' preferences. This can however only be done by acknowledging the limited scope of validity for welfare analysis.

A first approach I count as behaviourist is the 'Unified framework' of Bernheim (2016), who considers the possibility of 'true preferences' while explicitly rejecting its general use:

> As an example, consider decisions involving ordered lists of options, such as candidates enumerated on a ballot. Various studies document a tendency to pick the alternatives listed first … Here, the ordering defines the decision frame, and all orderings potentially distort the expression of "true preferences." At first this may appear contrary to what I have written above, but the explanation is simple: as formulated, *the theory only pertains to the limited choice domain within which the distortion occurs*. (pp42-43, emphasis added)

Bernheim presents the 'two core tasks of the Unified Framework' as follows:

> **Step 1**: we identify all decisions that merit deference (the *welfare-relevant domain*)
> **Step 2**: we construct a welfare criterion based (at least in part) on the properties of choices within that domain (p.43)

The difficulty lies then in defining the correct process by which we could identify the 'welfare-relevant domain'. An issue with Bernheim's initial contribution to BWE (Bernheim and Rangel, 2007, 2009) was indeed that the identification of the 'suspect' generalised choice situations seemed to require the exact same kind of reasoning as in the account based on the inner rational agent (and Bernheim himself acknowledges that his position evolved in the last years – Bernheim 2016,

[14] It is indeed striking to note that, despite the 'as judged by themselves' clause, behavioural paternalists seem to share the view that *all* individuals should agree on a wide range of preferences, such as saving more for their retirement, eating less calories, adopting healthy lifestyles, etc. (Lecouteux 2016, pp.194-195). While those are reasonable claims about what would count as a 'good life', the claim that all agents would behave like that, *if freed from reasoning imperfections*, is a mere postulation, and is only supported by casual psychological arguments.

p.13). Here the criterion is that the foundation for a specific behavioural model is 'compelling' (Bernheim 2016, p.46), which may be a suitable solution for specific cases, but remain silent for too many problematic situations. Just as the welfarist approach, the risk is that the economist' own judgement will determine the cases 'that merit deference'. As I will argue later, a criterion that I think could satisfy Bernheim's goal is to restrict welfare analysis to situations labelled by Savage (1954, p.86) as *microcosms* – which are however quite rare in practice.

A second approach that intends to keep choices as the primitive of welfare analysis, while acknowledging the possibility of incoherent preferences, is the 'quantitative intentional stance' [QIS] of Harrison and Ross (2018). The general philosophical framework of their approach is Dennett's (1987) *externalist* account of preferences and beliefs. Rather than considering that preferences and beliefs are inner mental states that cause the individual's behaviours, they are defined as attributions to ourselves and others that make our behaviours socially understandable. Taking the intentional stance toward an agent:

> 'consists in assuming that the agent's behavior is guided by goals and is sensitive to information about means to the goals, and about the relative probabilities of achieving the goals given available means' (Harrison and Ross 2018, p.20)

There is however more than a mere instrumentalist account of the individual's intentionality, because the individual herself must take the intentional stance toward herself, to 'try to make all of [her] material cohere into a single good story' (Dennett 1992, p.114), the person's 'autobiography' (Dennett 1994, p.74). This definition of the self in terms of narrativity presupposes the existence of 'a context of *discourse*, where a settled semantics is operative and there exists an audience interpreting the events in question' (Christman 2009, p.81). More than a useful tool for the theoretician to represent peoples' behaviours, preferences and beliefs constitute our *shared language* required in the process of socialisation.

From this perspective, looking for a notion of welfare does not require investigating the mental states of the agent, but simply interpreting the agents' behavioural and cognitive ecologies in terms of the economist's language of subjective expected utility. However, while genuine 'real patterns' (Dennett 1991) in choices are captured through the attribution of preferences and beliefs, choices also involve some 'noise', which are not necessarily seen as cognitive deficiencies on the agent's behalf, but rather as measurement limitations of the theoretician's model. The lab then offers a 'clean' environment within which the choices of the agents can be interpreted as the satisfaction of preferences given subjective beliefs. Harrison and Ng (2016, 2018) and Harrison and Ross (2018) for instance characterise the risk preferences of agents by eliciting the most likely preference structure (expected utility or rank-dependent utility) in simple experimental tasks, and use those risk preferences as the welfare metric for choices among insurance products or portfolios. In the absence of an independent measure of risk preferences, Alekseev *et al* (2018) develop a method to estimate and disentangle econometrically the 'noise' and associated 'welfare' – as the most likely characterisation of preferences and stochastic term in a stochastic choice model – and offer a direct way to measure the welfare costs of noisy behaviours.

Unlike Bernheim's unified framework that maintains a choice-based approach to welfare economics but is limited by a loose criterion of Pareto indifference, the methodology of the QIS is able to develop operational measures of welfare. The cost is however – just as with Bernheim – a restricted scope of validity for BWE, which can deal with 'preferences that violate EUT but are nevertheless well-ordered' (Harrison and Ross 2018, p.22). The typical cases for which BWE can be applied require for instance the existence of certainty equivalents, which is sensible for e.g. insurance choices, but which may be more problematic for other contexts dealing with non-monetary outcomes, such as health or education. I will argue below that the QIS constitutes an adequate framework for choice situations characterised by Savage as *pseudo-microcosms* (and more generally, 'not-too-uncertain' large worlds).

2.3 Constitutional approaches

An alternative – and radically different – way to approach the reconciliation problem is to reject preference satisfaction as a valid normative criterion, and to shift normative appraisal from 'welfare' (whatever it means) to either the environment of choice – the *constitutional* approach – or the process of decision-making – the *procedural* approach.

According to the constitutional approach, (B) remains a tautology, though it is acknowledged that the preferences representing peoples' choices may be non-standard. The main (though not only) proposal endorsing this perspective is Sugden's *contractarianism*. Following Buchanan (1986), Sugden (2013) complains that neoclassical and behavioural welfare economics are usually addressed to a benevolent despot, who is in charge of the optimal organisation of social relations, and acts as a philanthropist towards the rest of the society. Sugden argues on the contrary for a contractarian approach to normative economics, according to which the theoretician is not speculating about what is good for the population, but is rather directly addressing citizens, advising them how to reach mutually beneficial agreements. The contractarian approach shifts normative appraisal from welfare (what the individual obtains *in fine*) to opportunities (the set of alternatives from which the individual can choose). Rather than designing policies that 'give' to individuals what we think maximises their welfare, our goal as normative economists is to ensure that individuals have the opportunity to satisfy any preferences that they *might have* (the fact that their preferences are consistent or not is of no relevance for us). Agents are considered to be better or worse off depending on their opportunity sets: this criterion allows Sugden (2004, 2007) to reformulate the first fundamental theorem of welfare economics in terms of the 'opportunity criterion' rather than Pareto optimality. The crux of his argument is that the equilibrium of competitive markets is characterised by a maximisation of opportunity sets for the agents (maximisation in the sense that increasing further the set of one agent would require decreasing the set of another), which is a result of market properties rather than agents' preferences. Under the *additional assumption* that preferences are consistent and context-independent, the maximisation of opportunity sets implies Pareto optimality (meaning that Pareto optimality is only a fortunate by-product of preference consistency, while the normative appeal of markets lies in the maximisation of opportunity sets).

According to Sugden, what matters are specific properties of the choice environment, rather than the mental states of the individual. He acknowledges that in certain cases of *self-acknowledged* failures of self-control (e.g. heroin addicts), letting the individual makes her own choices might be normatively problematic, though he considers that such cases of genuine problems of self-control are quite rare (Sugden 2017). Unlike proponents of nudging who see self-control problems in most of our behaviours (with enduring conflicts between our short-term and long-term selves), a person is considered as a continuing *locus of responsibility*, treating her past and future actions as her own, whether or not those actions were or will be what she would like them to be now (Sugden 2004, p.1018). This quality of responsible personhood gives a normative force to the judgement of the agent herself on her own actions – and the theoretician, while legitimately modelling her choices with non-standard preferences, is not entitled to formulate any value judgement on those preferences.

In parallel to Sugden's contractarianism, other contributors have also (with a slightly different perspective) focused on the properties of the choice environment rather than choice outcomes, by emphasising the role of *democratic* governance in discussions around BWE. BWE – and more specifically the nudging agenda – is identified as a *technocratic* attempt to steer people's behaviours in certain directions, with little legal or democratic control. The generalisation of nudging raises questions about its legality (van Aaken 2019) and its institutional consequences with a primacy of current scientific knowledge over public deliberation in the design of public policies (Lepenies and Malecka 2015, 2019). Furthermore, explicitly representing the interaction between citizens and bureaucrats in charge of designing nudges suggests that nudges may lead to the reinforcement of current social norms – which leads to conservative policies which are far from certain to be welfare-enhancing, and could conflict with the constitutional interests of citizens (Schnellenbach 2012, 2016, Schubert 2017). From this perspective, preference inconsistencies do not constitute a fundamental problem for normative analysis, though their exploitation by nudges may jeopardise the basic principles of liberal democracies. Advocating a form of paternalism that recognises two classes of citizens (the 'rational' and the 'irrational'), is seen as deeply concerning, as when Camerer *et al* (2003) call 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)

The constitutional approaches maintain the tradition of mindless welfare economics which values the freedom of choice of the agent (and more accurately, of the citizen), knowing that such freedom can only be properly exercised if we are embedded in an institutional context whose aim is to promote individual (economic and political) freedom.

2.4 Procedural approaches

A last alternative approach to the reconciliation problem is to keep the tradition of mindful welfare economics by investigating the actual process of decision-making, while shifting normative appraisal from the outcome to the process of decision-making. Unlike constitutional approaches that focus on the institutional context within which the agents choose, the focus here is on the factors that directly shape the preferences of the individual.

A first line of argument, initially advanced by Sunstein (1991) – before he developed a more welfarist position in his later works on libertarian paternalism – is to locate normative appraisal on the process of *preference formation* rather than *preference satisfaction*, which means discussing the conditions for enhancing the person's *autonomy* rather than *welfare*. This idea has been pursued in different directions, with a common starting point in the position that preferences are evolving and cannot be considered as given. We can find this type of argument in Binder and Lades (2015) defence of 'autonomy-enhancing paternalism' (in particular pp.13-15), or Buchanan's (1979) notion of 'creative choice' developed by Schubert (2015), Dold (2018), and Dold and Schubert (2018). Those contributions emphasise the role of learning and individual development in the definition of welfare, and advance a significant role for education policies aiming to guide individuals to form their own preferences in an autonomous way (Lecouteux 2015b, pp.132-133).

The most elaborate and developed form of such 'educating' policies is the *boost* program put forward by Grüne-Yanoff and Hertwig (2016, 2017), which is directly inspired by the simple heuristics program of Gigerenzer *et al* (1999).[15] Instead of strategically harnessing the agents' bounded rationality with nudges, as a means to maximise their welfare, *boosts* intend to foster individual competences by training decision makers to employ more adaptive heuristics. Unlike nudging, boosting requires the active involvement of the agent, and seems thus less paternalistic – the archetypical illustrations of the two policies are indeed default options for nudging (which are effective as long as the agent remains passive) and teaching people to reframe probabilities as natural frequencies for boosting (which requires the active participation of the agent). By enhancing the cognitive processes of the agent, a boost will not be specific to one situation, and may generate positive spillovers for the agent – who may now also resist manipulation attempts by e.g. marketers. This suggests that boosts may impact behaviours that are out of the reach of nudges[16] and also impact behaviours for a longer period of time after the intervention.[17]

[15] Unlike the heuristics & biases program pioneered by Kahneman and Tversky, which conceptualizes bounded rationality in terms of deviations from rational choice, the simple heuristics program emphasizes the importance of the adaptivity of behaviours in relation to the environment. See Ortmann and Spiliopoulos 2017 for a review.

[16] An illustration is improving how parents prepare and organize family meals, as a way to improve children's diet. See Dallacker *et al* 2018 for an investigation of the relationships between characteristics of family meals and children's dietary quality. The boost would consist here in training parents to adopt better practices, such as turning off television during meals, involving children in meal preparation, etc.

[17] The effectiveness of the intervention will however remain limited in time if the agent is also subject to other pressures aiming to reinforce the behaviours the boost intended to limit – such as the consumption of junk food by adolescents, which can be limited with an adequate boost (training people to systematically see advertisement campaigns as designed in the interest of big companies rather than consumers), but which is stimulated by a constant exposure to aggressive marketing (Bryan *et al* 2019).

While boosting seems *a priori* ethically less problematic than nudging, a significant difficulty is that boosts are not necessarily available, and behaviour change may take a much longer time than with nudges (investing in boosts 'entails significant upfront costs, while the nudge approach [is] viewed as a fertile source of low-cost means for inducing changes in behaviour', Earl 2018, p.121). In cases where the degree of 'teachability' of heuristics is low, nudges may be a more effective strategy (Grüne-Yanoff *et al* 2018, p.257). Furthermore, boosting requires discovering effective heuristics, which is very far from trivial (Lieder 2018), and may also impose arbitrary judgements about what counts as an 'effective' heuristic (Sims and Müller 2019).

3. Reframing the reconciliation problem: welfare economics in large worlds

The classification above highlights the heterogeneity of approaches seeking to address the reconciliation problem, which seem at first sight hard to make compatible. The risk is then to start an endless debate amongst antagonistic positions, with no clear guide about how to choose public policies in different contexts. I will argue however that Savage's distinction between small and large worlds offers a simple framework to clarify the debate, and to determine which policies to be applied depending on the choice problem. My argument is that the four approaches discussed above attack the exact same problem – setting the foundations of normative economics in a large world – and that they offer complementary solutions, depending on the nature of the uncertainty of the choice problem (which determines whether it makes sense or not to maintain (B) and (N)).

3.1 Small and large worlds

The distinction between small and large worlds was introduced by Savage to delimitate the scope of validity of his Bayesian representation of choices. With the notable exception of Binmore (2009), this distinction has been largely overlooked in the literature in decision theory, probably because of its extremely vague definition. Savage explains the difference between small and large worlds as follows:

> 'The point of view under discussion may be symbolized by the proverb, "Look before you leap,", and the one which it is opposed by the proverb, "You can cross that bridge when you come to it." When two proverbs conflict in this way, it is proverbially true that there is some truth in both of them, but rarely, if ever, can their common truth be captured by a single pat proverb. One must indeed look before he leaps, in so far as the looking is not unreasonably time-consuming and otherwise expensive; but there are innumerable bridges one cannot afford to cross, unless he happens to come to them.' (Savage 1954, p.16)

A choice problem in which 'Look before you leap' is a reasonable principle of choice, i.e. in which the individual can anticipate all the possible consequences and plan in advance all her future moves given the possible states of the world, 'in so far as the looking is not unreasonably time-consuming and otherwise expensive' is called a *small world*. Otherwise:

> 'Carried to its logical extreme, the "Look before you leap" principle demands that one envisage every conceivable policy for the government of his whole life (at least from now on) in its most minute details, in the light of the vast number of unknown states of the world, and decide here and now on one policy. This is utterly ridiculous, not – as some might think – because there might be later cause for regret, if things did not turn out as had been anticipated, but because the task implied in making such a decision is not even remotely resembled by human possibility' (Savage 1954, p.16)

In such *large worlds*, Savage proposes that agents will successively solve tractable problems by focusing on 'isolated decision situations'. For instance, if you are asked to meet a stranger in Paris, you will first frame the problem as 'find a focal point in Paris' and probably go to the Eiffel Tower. But once there, you will realise that the Eiffel Tower is not really a meeting 'point', and will then frame the next problem as 'find a focal point at the Eiffel tower', and then go to e.g. the ticket office, etc. Indeed, according to Savage, 'to cross one's bridges when one comes to them means to attack relatively simple problems of decision by artificially confining attention to so small a world that the "Look before you leap" principle can be applied there' (Savage 1954, p.16). He however confesses being 'unable to formulate criteria for selecting these small worlds', while believing 'that their selection may be a matter of judgment and experience about which it is impossible to enunciate complete and sharply defined general principles' (p.16).[18]

Note that the distinction above allows us to define more precisely the idea of 'context' largely used in BWE. While a large world has a countless number of states of the world (an arbitrary precise description of the situation, such as e.g. the exact temperature in Centigrade degrees), the states of the world in the small world constitute a *partition* of the states of the large world (e.g. whether the temperature is below or above 0°C) – see Larrouy and Lecouteux 2018 for a detailed presentation. Given a small world representation of the initial problem (i.e. how the agent partitions the set of states of the large world), Savage shows that when the choices respect certain rationality postulates, it is as if the agent was maximizing her subjective expected utility, with the utility function defined over the states of the small world. In this framework, the 'context' can be defined as the variations in the states of the large world (which can be perceived by the theoretician) that do not affect the utility in the large world – this set of states of the large world constitutes an event in the large world, which can also be defined as a single state in a smaller world. This means that what counts as the context depends on 'how *we*, the theorists, *"*cut up the

[18] Larrouy and Lecouteux (2018) argue that Bacharach's variable frame theory can offer an adequate framework to model how agents select their small world representations, and suggest that the process driving the convergence of individual small-world representations to *common* social representations is akin to mindshaping (Zawidski 2013). Ross (2005, chapter 7) – echoing Binmore's (1994, 1998) evolutionary analysis of the emergence of norms and game representations in coordination problems – describes a similar dynamic process, in which the convergence to equilibrium strategies in coordination games is made possible thanks to the emergence of public narratives that will align the agents' expectations. Narratives – and then, the agents' preferences and representations of the game – depend on the evolving narratives of the other agents and are thus socially constructed.

world"' (Bacharach 2006, p.13), i.e. the context is defined by the economist's *own views* on which partition of the states of the world is relevant to the choice problem.

3.2 Large worlds and the reconciliation problem

Savage characterises the idea of an isolated decision situation with the notion of (pseudo-)microcosm. A small world representation of the initial problem is called a *microcosm* if and only if the agent would exhibit the same subjective probability judgments and utility functions had he chosen consistently based on the states of the large world or based on the states of the small world in question. This means that when a small world is a microcosm, it is as if the agent had unlimited cognitive abilities and was able to choose consistently in the large world. On the other hand, when the agent respects the rationality postulates given her small world representation, but the induced utility function and subjective beliefs are different from the ones that would be induced by a consistent chooser in the large world, we have a *pseudo-microcosm*.

The notions of microcosm and pseudo-microcosm help us to reframe the reconciliation problem in terms of small world representations. If we assume that agents choose consistently given their small world representation, nothing guarantees that the preferences used by the theoretician to represent her choices are based on the *same* small world representation. Neoclassical economics supposes that the theoretician and the agent always share the same representation, and therefore restricts welfare analysis to *microcosms*. The strategy pursued by Bernheim is essentially the same, as his approach to define the welfare-relevant domain aims at avoiding as much as possible potential conflicts regarding what could count as mistakes: limiting oneself to microcosms, when the theoretician and the agent agrees on the small world representation, is thus a good criterion to define the welfare-relevant domain.

If we consider on the other hand that the representation of the agent differs from the representation of the theoretician, the preferences revealed in the agent's choices will be non-standard.[19] That is, the utility function representing the choice of the agent is different from the function that the theoretician would have used, if put in the situation of choice instead of the agent. If we have good reasons to believe that the theoretician is better informed than the agent about the choice problem (and therefore use a finer partition of the large world than the agent), then we can interpret this situation as a pseudo-microcosm. This is the line of argument of welfarist approaches, in which the true preferences are supposed to be consistent with the representation of the theoretician, while the revealed preferences of the agent may deviate because of an *inadequate* representation of the choice problem (e.g. treating elements of the context as relevant for the problem). The QIS, on the other hand, presupposes that the small world representation of the agent

[19] In the case of retirement savings for instance, the theoretician may be tempted to frame the problem as discounting future outcomes with respect to time (for which exponential discounting is normatively compelling) while an agent who discounts her future outcomes with respect to psychological connectedness (i.e. her degree of 'closeness' with her future selves, see Parfit 1984) would exhibit hyperbolic discounting in the theoretician's small world representation (Lecouteux 2015a).

is the correct one, and tries to infer econometrically this representation – while still allowing the expression of some noise in the behaviour of the agent when embedded in a large world. This is why, as in Harrison and Ross (2018), the lab offers a good environment (it is indeed a small world by construction) to first elicit the risk and time preferences of the agents, before recovering the preferences of the agent in a larger world.

Note that the discussion on small worlds and pseudo-microcosms requires assuming that the actual choice problem can indeed be represented by a large world, with arbitrarily precise states of the worlds (Savage names this 'largest' world – with an arbitrarily fine partition – the 'grand world'). Savage himself is however skeptical of the idea (mentioning his reluctance at pages 83 and 90 in the *Foundations*), and uses this approach of defining small worlds as partition of the grand world 'for want of a better one' (p.83). In particular, considering situations of choices with a more fundamental uncertainty, as the property of the mind-environment system (Todd and Gigerenzer 2012, p.18), can seriously undermine the plausibility of the counterfactual existence of coherent choices in such environments. Such situations may call for alternative solutions that do not rely on the standard Bayesian tools of preferences and subjective beliefs (i.e. the approaches that reject (N) and the criterion of preference satisfaction), simply because it is not technically possible to define sufficiently precise states of the world.

3.3 Bounded rationality in small worlds

Welfare economists have traditionally assumed working in microcosms, considering that models (such as a payoff matrix to represent a specific game) are idealisations of real-life situations, while assuming that the agents share the exact same representation of the choice problem. The difficulty is that real individuals almost never interact in such small worlds (except during lab experiments), and therefore arrive in the lab with the tools they have developed to interact in actual large worlds, which might be deeply inadequate in this context.[20] Vernon Smith expressed this concern as follows (in an unpublished letter to Harsanyi, 1989):[21]

> 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 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

[20] On the other hand, if allowed to play in the lab for a sufficiently long period of time, the agent may adapt her behaviour and *in fine* behaves in accordance with the prescriptions of neoclassical economics (Binmore 1999). This however only means that the agent has some faculty of adaptation, and not that she would also choose in large worlds following the prescriptions of neoclassical economics if given enough time to adjust to the environment.

[21] I am very grateful to Dorian Jullien for sending me a scan of this letter during his stay at Duke University.

> "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.

Viewed in this light, the deviations from expected utility theory do not reveal mistakes on the part of the agent: she may act in a perfectly sensible way, while having in mind a significantly different framing of the problem than the theoretician's one. In the words of Sugden (2018), 'it is as if decision-makers are held to be at fault for failing to behave as the received theory predicts, rather than theory being at fault for making incorrect predictions.'

It is important to note here the specific place of welfarist approaches compared to the others when coming to the interpretation of lab experiments. The model of the inner rational agent presupposes that agents are pathologically maladjusted to their environment, and are predictably *stupid* (rather than 'predictably irrational'). Lab experiments seem indeed to highlight that people choose very poorly in small world environments, which – given the 'simplicity' of such situations compared to more uncertain decision problems – is interpreted as evidence of severe cognitive deficiencies. It is quite telling that the archetypal 'Human' for Sunstein and Thaler is Homer Simpson! Welfarist approaches picture the individual as a defective Bayesian agent embedded in small worlds only – while most (if not all) decisions faced by agents which are relevant for welfare economics take place in large worlds. Inferring any policy recommendation from BWE thus requires an analysis of how people choose in such large worlds, and cannot be limited to the use of a catalogue of biases that are exhibited in environments within which the agents almost never act. This is a reason motivating Gigerenzer's analysis of simple heuristics rather than optimisation strategies, 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). The normative relevance of a strategy – optimisation or simple heuristics – depends on the nature of the environment, with a preference for simple heuristics when discussing choices in large worlds (Gigerenzer and Sturm 2012, pp.262-264).

4. Welfare and public policies in large worlds

The notion of economic welfare used both in neoclassical and behavioural welfare economics is mostly suitable for small worlds, simply because inferring preferences and beliefs when embedded in a large world may not be possible. The question that must now be answered is what normative criterion (and then, which policies) should be applied when looking at large worlds. My suggestion is that the characteristics of the choice situation – whether the uncertainty is measurable or not, and whether we, as theoreticians, are in a higher epistemic position than the agents we intend to model[22] – will indicate whether we can maintain statements (B) and (N), and then point out to the adequate approach among the four solutions discussed previously.

[22] By 'higher epistemic position', I mean that we, as theoreticians, can expect to know more than the agents about (i) the choice situation and/or (ii) the agents' characteristics. Distinguishing between sources of uncertainty related to either the environment or the agent is indeed common practice – as well as psychologically meaningful, see Fox

Consider first a choice situation in which uncertainty is measurable by probabilities. In such (not so uncertain) environments, it is common to assume that we know at least as much as the agents about the environment.[23] If, furthermore, we have good reasons to believe that we have a better knowledge than the agent herself about her *own preferences*, then a possible objective would be to guarantee – by the imposition of adequate incentives or nudges – that the agent manages to choose whatever maximise her true preferences. This is the methodology of the welfarist approach, which endows the theoretician with a superior knowledge both over the subject and environment of choice. Unlike the conventional wisdom of BWE, I would suggest however that such cases are empirically rather limited, and mostly correspond to lab experiments, and self-acknowledged cases of self-control failures. The approach I would privilege when considering cases with a measurable uncertainty is to assume that the preferences and beliefs of the agent are *a priori* unknown to us, so we are in an impoverished epistemic position relative to that of the agent we model. When observing the choice of the agent, we estimate her preferences and beliefs by assuming that she behaved as though seeking to maximise her subjective expected utility, given her small world representation. This offers us a measure of economic welfare for this specific choice problem, while possibly allowing for noise in the transcription of the welfare function – this is precisely the approach of the QIS. Note that once we have a welfare measure, it is *possible* to use preference satisfaction as a normative criterion (and keep (N)), but we may also consider – for ethical reasons – that the relevant normative criterion is not preference satisfaction (such as opportunity for Sugden).

Consider now choices with non-measurable uncertainty (e.g. Knightian uncertainty). We therefore consider large worlds for which it does not make sense to refer to the grand-world from which the agent could build a small-world representation of the problem. We cannot therefore infer preferences and beliefs from the choices of the agent, and thus will lack an operational measure of welfare. As suggested by Savage himself, an option to choose under complete ignorance is to use *heuristics* (see Binmore 2009, p.156): this proposal is in line with Simon's (1955, 1956) project of developing a psychologically realistic theory of rational choice for actual agents under a situation of radical uncertainty. Rather than looking for an 'optimal' choice that would be cognitively too costly, the criterion of (bounded) rationality is that the agent reaches a sufficiently satisfying outcome. What matters is therefore *ecological rationality*, i.e. 'the fit of a strategy to the structure of an environment' (Hertwig *et al*, 2019, 367). If we have good reasons to believe that we know more effective heuristics than the ones the agent uses – being in this specific regard in a higher epistemic position – then boosting seems an appropriate policy. If, however, we cannot claim our higher epistemic position (because for instance of a lack of tacit knowledge that is specific to an organisation), and are unable to advise more fitted heuristics for the agents, our role should be limited to guaranteeing a well-functioning environment – i.e. endorse a constitutional position.

To summarise, when the uncertainty is measurable (we consider not-too-uncertain environments which can reasonably be represented by small worlds), we can apply the expected

---

and Ülkümen 2011 – in particular when the uncertainty is measurable with probabilities. If it is not measurable, the uncertainty will be about the degree of fit between the agent and the environment (Kozyreva *et al* 2019).

[23] It is for instance reasonable – given the very nature of our profession – to assume that we are at least as expert as the agents we advise when considering financial choices or market decisions.

utility framework, and can legitimately keep (N), as is common practice among welfare economists. When facing a more radical uncertainty however, preference satisfaction ceases to be a potential normative criterion, and we should reject (N). Furthermore, in cases where we have good reasons to expect to know more than the agent about the choice problem (about herself and/or the environment), then we can expect to be able to improve the agents' choices (and reject (B)) by either adequate incentives/nudge or adequate training/boosts. If, on the other hand, we consider that we are in a similar or impoverished epistemic position relative to that of the agent, then we should respect the agents' choices and keep (B).

The different solutions to the reconciliation problems reviewed in section 2 are therefore *complementary* approaches, and their adequate domain of application depends on the nature of the choice problem:

	Lower epistemic position (keep (B))	Higher epistemic position (reject (B))
Measurable uncertainty (keep (N))	Behaviourist (QIS)	Welfarist
Non-measurable uncertainty (reject (N))	Constitutional	Procedural

Concluding remarks

I have argued in this chapter that Savage's distinction between small and large worlds offers an adequate framework to address the reconciliation problem, and can more pragmatically offer a guide to identifying the best policy alternatives depending on the nature of the uncertainty of the choice problem. This however requires economists to be aware of their own epistemic position. While we may be tempted (and this is a perfectly reasonable position in many cases) to attribute to ourselves higher knowledge as 'experts' – it is no surprise that the approaches that have generated the greatest enthusiasm in the literature are nudges and boosts – we should take care not to *systematically* presuppose our higher expertise. We should always keep in mind that empirical deviations from our theoretical recommendations may either be the product of genuine errors from the agent (due to a decision bias or the use of an ineffective heuristic) or of measurement errors from our imperfect models.

The greatest challenge that remains is to formulate policy prescriptions when looking at very uncertain environments, within which it seems practically impossible to define operational measures of individual welfare. The reason is probably that looking at a theory of choice in large worlds requires rejecting the *atomistic* conception of the individual, and consider agents as embedded in a socio-historical environment. Remarkably, apart from the welfarist approach that identifies the agent with her true preferences (i.e. her inner rational agent – but see Davis (2011) on the self-referential problem of preferences and identity of the *homo economicus*), all the other approaches introduced above advance a *social* view of the self. The Dennettian background of the QIS is a clear rejection of methodological individualism, Sugden's contractarian proposal

conceives individuals as potential parties to mutual arrangements (i.e. to a social contract), and the boost agenda is aimed at increasing the *ecological* rationality of the agent's heuristics. Any notion of economic welfare suitable for large worlds should thus integrate a normative assessment of the environment, and of the relationship between the agent and her environment. As long as we lack an unambiguous measure of welfare as the degree of fit between the agent and the environment – which might be necessary to properly assess the effectiveness of heuristics proposed by the boost program – the best and possibly only alternative is to directly form normative judgements about the constitution of the environment (just as the constitutional approaches recommend), which may however lead to arbitrary judgements about what counts as a good constitution.

References

Alekseev, A., Harrison, G. W., Lau, M., & Ross, D. (2018). Deciphering the Noise: The Welfare Costs of Noisy Behavior.

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

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

Bernheim, B. D., & Rangel, A. (2007). Toward choice-theoretic foundations for behavioral welfare economics. *American Economic Review*, *97*(2), 464-470.

Bernheim, B. D., & Rangel, A. (2009). Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics*, *124*(1), 51-104.

Binder, M., & Lades, L. K. (2015). Autonomy-Enhancing Paternalism. *Kyklos*, *68*(1), 3-27.

Binmore, K. (1994). *Game Theory and the Social Contract. Playing Fair* (Vol. 1). MIT press.

Binmore, K. (1998). *Game Theory and the Social Contract: Just Playing* (Vol. 2). MIT press.

Binmore, K. (1999). "Why experiment in economics?." *The Economic Journal* 109: 16-24.

Binmore, K. (2007). Rational decisions in large worlds. *Annales d'Economie et de Statistique*, 25-41.

Binmore, K. (2009) *Rational decisions*. Princeton University Press.

Bryan, C. J., Yeager, D. S., & Hinojosa, C. P. (2019). A values-alignment intervention protects adolescents from the effects of food marketing. *Nature human behaviour*, *3*(6), 596-603.

Buchanan, J. M. (1979). "Natural and artifactual man." *What should economists do*: 93-112.

Buchanan, J. M. (1986). *Liberty, market and state: Political economy in the 1980s*. Wheatsheaf Books.

Camerer, C. (2008). The case for mindful economics. *Foundations of positive and normative economics*, 43-69.

Camerer, C. (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.

Christman, J. (2009). *The politics of persons: Individual autonomy and socio-historical selves*. Cambridge University Press.

Dallacker, M., Hertwig, R., & Mata, J. (2018). The frequency of family meals and nutritional health in children: a meta-analysis. *Obesity Reviews*, *19*(5), 638-653.

Davis, J. B. (2010). *Individuals and identity in economics*. Cambridge University Press.

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Dennett, D. C. (1988). Why everyone is a novelist. *TLS-THE TIMES LITERARY SUPPLEMENT*, (4459), 1016.

Dennett, D. C. (1991). Real patterns. *The journal of Philosophy*, *88*(1), 27-51.

Dennett, D. C. (1992). The self as a center of narrative gravity. In *Self and consciousness: multiple perspectives*, ed. F. Kessel, P. Cole, and D. Johnson. Hillsdale, NJ: Erlbaum.

Diener, E., & Biswas-Diener, R. (2011). *Happiness: Unlocking the mysteries of psychological wealth*. John Wiley & Sons.

Dold, M. F. (2018). Back to Buchanan? Explorations of welfare and subjectivism in behavioral economics. *Journal of Economic Methodology*, *25*(2), 160-178.

Dold, M. F., & Schubert, C. (2018). Toward a behavioral foundation of normative economics. *Review of Behavioral Economics*, *5*(3-4), 221-241.

Earl, Peter E. "Richard H. Thaler: A Nobel Prize for Behavioural Economics." *Review of Political Economy* 30.2 (2018): 107-125.

Frijters, P., Clark, A. E., Krekel, C., & Layard, R. (2020). A happy choice: wellbeing as the goal of government. *Behavioural Public Policy*, *4*(2), 126-165.

Fox, Craig R., and Gülden Ülkümen. "Distinguishing two dimensions of uncertainty." *Perspectives on thinking, judging, and decision making* (2011): 21-35.

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.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Graaff, J. D. V. (1967). *Theoretical welfare economics*. CUP

Griffin, J. (1986). *Well-being: Its meaning, measurement and moral importance*.

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory?. *Minds and Machines*, *26*(1-2), 149-183.

Grüne-Yanoff, T., Marchionni, C., & Feufel, M. A. (2018). Toward a framework for selecting behavioural policies: How to choose between boosts and nudges. *Economics & Philosophy*, *34*(2), 243-266.

Guala, F. (2019). Preferences: neither behavioural nor mental. *Economics & Philosophy*, *35*(3), 383-401.

Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. *The foundations of positive and normative economics: A handbook*, *1*, 3-42.

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

Harrison, G. W., & Ng, J. M. (2016). Evaluating the expected welfare gain from insurance. *Journal of Risk and Insurance*, *83*(1), 91-120.

Harrison, G. W., & Ng, J. M. (2018). Welfare effects of insurance contract non-performance. *The Geneva Risk and Insurance Review*, *43*(1), 39-76.

Harrison, G. W., & Ross, D. (2018). Varieties of paternalism and the heterogeneity of utility structures. *Journal of Economic Methodology*, *25*(1), 42-67.

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

Hausman, D. M. (2016). On the econ within. *Journal of Economic Methodology*, *23*(1), 26-32.

Hausman, D. M. (2020). Enhancing welfare without a theory of welfare. *Behavioural Public Policy*, 1-16. doi:10.1017/bpp.2019.34

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973-986.

Hertwig, Ralph, Timothy J. Pleskac, and Thorsten Pachur. (2019) *Taming uncertainty*. MIT Press.

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.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The quarterly journal of economics*, *112*(2), 375-406.

Kozyreva, A., T. Pleskac, T. Pachur, & R. Hertwig (2019). Interpreting Uncertainty: A Brief History of Not Knowing. In Hertwig *et al* (Eds), *Taming uncertainty*, 343-362.

Larrouy, L., & Lecouteux, G.. (2018) "Choosing in a Large World: The Role of Focal Points as a Mindshaping Device." GREDEG Working Paper.

Layard, R. (2011). *Happiness: Lessons from a new science*. Penguin UK.

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

Lecouteux, G. (2015b). *Reconciling normative and behavioural economics* (Doctoral dissertation, École Polytechnique).

Lecouteux, G. (2016). From homo economicus to homo psychologicus: The paretian foundations of behavioural paternalism. *Œconomia. History, Methodology, Philosophy*, (6-2), 175-200.

Lepenies, R., & Małecka, M. (2015). The institutional consequences of nudging–nudges, politics, and the law. *Review of Philosophy and Psychology*, *6*(3), 427-437.

Lepenies, R., & Małecka, M. (2019). Behaviour change: extralegal, apolitical, scientistic?. In *Handbook of behavioural change and public policy*. Edward Elgar Publishing.

Lieder, F. (2018). *Beyond bounded rationality: Reverse-engineering and enhancing human intelligence* (Doctoral dissertation, UC Berkeley).

McQuillin, B., & Sugden, R. (2012). Reconciling normative and behavioural economics: the problems to be solved. *Social Choice and Welfare*, *38*(4), 553-567.

Nussbaum, M. C. (2011). *Creating capabilities*. Harvard University Press.

Ortmann, A., & Spiliopoulos, L. (2017). The beauty of simplicity?(Simple) Heuristics and the opportunities yet to be realized. In *Handbook of Behavioural Economics and Smart Decision-Making*. Edward Elgar Publishing.

Robinson, J. (1979), Collected Works, Vol. 5, Oxford: Basil Blackwell.

Robinson, J. (1974 [1962]). *Economic philosophy*. Penguin Books.

Ross, D. (2005). *Economic theory and cognitive science: Microexplanation*. MIT press.

Savage, L. J. (1954). *The foundations of statistics*. Courier Corporation.

Schnellenbach, J. (2012). Nudges and norms: On the political economy of soft paternalism. *European Journal of Political Economy*, *28*(2), 266-277.

Schnellenbach, J. (2016). A conRossstitutional economics perspective on soft paternalism. *Kyklos*, *69*(1), 135-156.

Schubert, C. (2017). Exploring the (behavioural) political economy of nudging. *Journal of Institutional Economics*, *13*(3), 499-522.

Simon, Herbert A. "A behavioral model of rational choice." *The quarterly journal of economics* 69.1 (1955): 99-118.

Simon, Herbert A. "Rational choice and the structure of the environment." *Psychological review* 63.2 (1956): 129.

Sims, Andrew, and Thomas Michael Müller. "Nudge versus boost: A distinction without a normative difference." *Economics & Philosophy* 35.2 (2019): 195-222.

Singh, R., & Alexandrova, A. (2020). Happiness economics as technocracy. *Behavioural Public Policy*, *4*(2), 236-244.

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 (2004) The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. Am Econ Rev 94:1014–1033

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 (2007) The value of opportunities over time when preferences are unstable. Soc Choice Welf 29:665–682

Sugden, R. (2013). The behavioural economist and the social planner: to whom should behavioural welfare economics be addressed?. *Inquiry*, *56*(5), 519-538.

Sugden, R. (2017). Do people really want to be nudged towards healthy lifestyles?. *International Review of Economics*, *64*(2), 113-123.

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

Sunstein, C. R. (1991). Preferences and politics. *Philosophy & Public Affairs*, 3-34.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American economic review*, *93*(2), 175-179.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Tiberius, V., & Plakias, A. (2010). Well-being. In J. Doris & the Moral Psychology Research Group (Eds.), The moral psychology handbook (pp. 401–431). Oxford: Oxford University Press.

Todd, Peter M., and Gerd Ed Gigerenzer. *Ecological rationality: Intelligence in the world*. Oxford University Press, 2012.

Tversky, A., & Kahneman, D. (Eds.). (2000). *Choices, values, and frames*. Cambridge University Press.

van Aaken, A. (2019). Constitutional limits to regulation-by-nudging. In *Handbook of Behavioural Change and Public Policy*. Edward Elgar Publishing.

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.