Reconciling normative and behavioural economics: the problem that cannot be solved

Guilhem Lecouteux

Reconciling normative and behavioural economics: the problem that cannot be solved

Forthcoming in S. Badiei & A. Grivaux (Eds), *The Positive and the Normative in Economic Thought*, Routledge.

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

Acknowledgements

An earlier version of this work has been presented at the conference 'The Positive and the Normative in Economic Thought' (Dec. 2020). I am grateful to the participants for their comments and suggestions, to the conference organisers and guest editors Sina Badiei and Agnès Grivaux, as well as two anonymous referees. I also thank the audiences of the joint GREDEG-TRIANGLE seminar and of the conference 'The Soul of Economics' (Zurich, 2019) where preliminary versions of this work were presented.

Reconciling normative and behavioural economics: the problem that cannot be solved

Abstract (107 words): Behavioural economics has challenged the normative consensus that agents ought to choose following their own preferences. I argue that normative economists implicitly defended a criterion of the sovereignty of the autonomous consumer, and that current debates in normative behavioural economics arise from disagreements about the nature of the threats to autonomy that are highlighted by behavioural economics. I argue that those disagreements result from diverging ontological conceptions of the 'self' in the literature. I distinguish between the unitary, psychodynamic, and socio-historical conceptions of the self, and show how different positive theories about preferences and the nature of the agent may determine normative positions in normative behavioural economics.

Keywords: preference satisfaction, autonomy, welfare, reconciliation problem, socio-historical self

JEL Codes: B40, D02, D60, D91.

Word count: 8700

Introduction

In the introduction to a symposium on 'Reconciling normative and behavioural economics' published in 2012 in *Social Choice and Welfare*,[1] McQuillin and Sugden subtitled their paper 'Reconciling normative and behavioural economics: the problems to be solved.' Traditional welfare economics builds on the assumption that individuals have stable and context-independent preferences and uses preference satisfaction as a normative criterion. By calling this assumption into question, behavioural economics [BE] raises fundamental problems for normative economics: if people's preferences are likely to change over time, or to depend on

[1] The symposium originated from a conference on this theme held in 2008 at the University of East Anglia.

apparently irrelevant aspects of the choice situation, can we still form normative judgments about people's choices based on their revealed preferences? The *reconciliation problem* requires clarifying whether the traditional principle of consumer sovereignty (which I equate here with the preference satisfaction [PS] criterion) should be preserved even in the presence of apparently incoherent choices and behaviours.

Several approaches have been suggested in the literature to tackle the reconciliation problem, advancing different normative criteria that could be used in normative analysis. Most of the debates have focused on the question of how to infer an adequate welfare metric from the possibly incoherent choices of the individuals – I will designate this part of the literature as *behavioural welfare economics* [BWE], while keeping the expression *normative behavioural economics* [NBE] for the literature including both BWE and the contributions advocating non-welfarist criteria (see Harrison (2019) and Lecouteux (2021a) for recent overviews of the literature in NBE). The main strategy endorsed in BWE consists in treating departures from conventional rational choice theory as mistakes, and then uses the satisfaction of the 'true' preferences of the individuals – the preferences that they would have revealed, were they not hampered by psychological biases – as the normative criterion (see Infante *et al* 2016 for a critical review).

My aim in this chapter is to approach the reconciliation problem from a slightly different perspective: rather than directly proposing a normative criterion, I will investigate first the conditions under which PS might be normatively appealing. Once this point is clarified, it could become easier to identify the core reasons of the disagreements within the literature, since we would be able to identify what kind of normative issues are raised by BE, depending on our justification of the PS criterion. I suggest that the main condition supporting PS is that individuals must be *autonomous*, implying that economists are implicitly committed to a principle of the 'sovereignty of the autonomous consumer'. The reconciliation problem stems from the observation that authors contributing to NBE have diverging ontological conceptions of the 'self', and thus disagree on the standards to qualify as autonomous. I identify three broad conceptions of the self in NBE: the *unitary* self, the *socio-historical* self, and the *psychodynamic* self.[2] Depending on one's conception of the self, we may conclude that BE offers direct evidence

[2] This typology is freely inspired from Meyer's (2005) 'Five faces of selfhood', with the 'socio-historical' conception encompassing both her 'social' and 'relational' self, while letting aside the 'embodied' self. I am not

that the autonomy of the agent is violated (or not). Diverging views about the nature of agents lead to different normative positions in NBE.

I start by reviewing the normative justifications of PS and highlight the coexistence of a welfarist and a contractarian approach, with different perspectives on the reconciliation problem. I then reframe this debate in terms of Sunstein's (1991) distinction between the stages of preference formation and preference satisfaction. I locate 'autonomy' at the stage of preference formation and discuss whether BE findings reveal that individual's autonomy is likely to be violated. Using overeating as a guiding illustration, I contrast how different explanations of preferences can lead to radically different policy recommendations.

Why does preference satisfaction matter?

A semantic issue: what are preferences?

Economic analysis is generally separated into two different branches: positive economics seeks to describe how economic agents behave, while normative economics aims at evaluating economic outcomes, policies, and institutions. Both approaches traditionally attribute to economic agents complete and integrated preferences,[3] i.e. preferences that are non-stochastic, context-independent, and internally consistent. Consistency is usually defined by (at least) the weak axiom of revealed preferences,[4] which entails that the choices of the agents can be represented by a complete and transitive preference relation (Hausman 2012: 26). Context-independence means that individual preferences remain stable across different *contexts*.[5] Individual behaviour is then described by assuming that the agents behave as if seeking to satisfy their preferences, and economic outcomes are desirable to the extent that individual preferences are satisfied.

---

claiming that this typology is more relevant in general, but merely that I believe that it adequately represents the current practices in NBE.

[3] I use here Sugden's (2018) terminology. Complete and integrated preferences typically satisfy the formal conditions allowing a utility representation of the preferences.

[4] 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*.

[5] I will not discuss here how to properly define the 'context' which is a question that is too often neglected in BWE. I propose elsewhere (Lecouteux & Mitrouchev 2021, Lecouteux 2021a) a definition in terms of small world representations by the theoretician.

Although the concept of preferences is central in economics, it is worth noting that little is said about what preferences *actually* are (Hausman 2012: 11). The normative interpretation of PS thus remains unclear. Hausman (2012: 1-2) suggests four main interpretations of the word 'preferences' for English-speakers:

- *Enjoyment comparison*: saying that Anna prefers *x* to *y* means that Anna enjoys more *x* than *y*. Preferences as enjoyment comparison are typically a matter of taste, such as preferring the taste of coffee to the taste of tea.
- *Comparative evaluations*: saying that Bob prefers *x* to *y* means that Bob judges *x* as better than *y* in some regard (according to one specific or any relevant criterion). Bob may for instance prefer to drink his coffee without sugar because it is healthier (even if he prefers – in terms of enjoyment comparison – to drink his coffee with sugar).
- *Favouring*: if a political party defends a policy of 'national preference' in terms of employment, then a native has a higher chance of being hired than an immigrant *ceteris paribus*. A specific class of individuals is therefore preferred (favoured), but without reference to an enjoyment comparison or a comparative evaluation.
- *Choice ranking*: saying that Carla prefers *x* to *y* means that she will choose *x* if she has to choose between *x* and *y*. When a waiter asks Carla whether she prefers coffee or tea, he only wants to know her choice, and does not ask her to provide a ranking in terms of enjoyment comparison or comparative evaluation.

Saying that PS matters does not mean the same thing whether we are thinking of preferences in terms of enjoyment comparisons, comparative evaluations or choice ranking.[6] If we consider preferences as enjoyment comparisons, PS matters if and only if the *experience* of enjoyment – my 'well-being' – in line with Bentham's utilitarianism, is valuable for itself (e.g. Layard 2011). When considered in the sense of comparative evaluations, my preferences represent my own conception of what is valuable for myself: even though I have a reason to prefer my coffee with sugar (because I enjoy the taste), and also a reason to prefer my coffee without sugar (because it is healthier), my preference for one option over another depends solely on the weights *I* give to

[6] I will not consider the interpretation in terms of "favouring" which is of little interest for the present discussion.

each reason in my preferences. PS matters now only if maximising my subjective 'welfare' (which I will use as a distinct term from 'well-being') is the adequate normative criterion. Finally, if preferences are understood as choice ranking, then saying that PS matters only means that it is a good thing to be able to choose whatever I want to choose whenever I want to choose it. The interpretation of the criterion is not welfarist anymore and emphasises the agent's freedom of choice (and of satisfying any preferences that the agent might have).

The normative justification of PS depends on the relationship between preferences and choice: if preferences are understood as the *cause* of our choices, then saying that PS matters means that it is a good thing that agents are able to pursue their own ends – whether 'well-being' if we consider that maximising one's well-being is the main driver of the agent's behaviour, or 'welfare' if we want to encompass a greater set of motives. If, however, we understand preferences as the *representation* of our choices, i.e. a *post hoc* formal representation of our behaviours, then saying that preference satisfaction matters does not imply any value judgement on our motivations – what matters is just that my choices are not constrained. Those two broad justifications of PS as a normative criterion depend on whether we endorse a *mentalistic* or *behaviouristic* account of preferences (see Guala 2019, Thoma 2021). While the former – the welfarist justification of PS – emphasises the value of one's own ends, the latter – that I will label contractarian – emphasises the value of individual choice, with no reference to predetermined ends.

The welfarist solution to the reconciliation problem

A common feature of the welfarist justifications of the PS criterion is that individual choices reveal an underlying *function* of welfare/well-being, i.e. that each alternative can be characterised by a unique level of welfare. The 'utility function' which is revealed through the choice of the agent – and which exists if and only if the revealed preferences are complete and integrated – can therefore be equated with her 'welfare function'. Preferring x to y reveals that the utility (and hence, welfare or well-being) associated to x is higher than the utility associated to y.

If, however, the preferences revealed in the agent's choices are not integrated (e.g. non-transitive or unstable), then there exists necessarily pairs of options x and y such that x is

preferred to *y* while the welfare associated to *y* is higher.[7] This means that PS does not offer any more an unambiguous normative criterion, because letting the agent makes her own choices will occasionally cause harm to the agent. Contrary to the consumer sovereignty principle, it is considered that in some situations, the agents may not act in their best interests: there is indeed a discrepancy between the preferences corresponding to the welfare function of the agent – her *true preferences*,[8] i.e. the function that the agent *ought* to maximise – and her revealed preferences. Given however that the only information we have at our disposal is the revealed preferences of the agent, we may wonder how we could elicit the underlying true preferences, and more fundamentally, whether such true preferences exist at all.

The idea that people could exhibit 'irrational preferences' which would differ from their true preferences has been discussed in economics for many years, with e.g. Pareto (1909: ch.3, §1) who describes a process of trial-and-error leading to the discovery of one's preferences. Kahn (1935: 25) also discusses market imperfections resulting from the satisfaction of 'preferences which obtain no justification […] in actual enjoyment, and the thwarting of which causes no loss of satisfaction'. It is however with the development of experimental economics that the idea gained momentum, mostly with Kahneman and Tversky and the 'heuristics and biases' program (see Heukelom 2014 for a historical discussion). Harsanyi (2008 [1977]: 29) similarly discusses irrational preferences, which are 'attributable, e.g., to various limitations in people's information-processing ability', while at the same time stating that 'such deviations from [rational choice axioms] do not affect their usefulness as axioms of a *normative* theory of rational behavior' (emphasis in original). More recently, Sunstein and Thaler (2003) and Camerer *et al.* (2003) argued that this discrepancy between the rational preferences of the agent and her revealed (often 'irrational') preferences offers a new justification for paternalistic regulations. Sunstein and Thaler argue that paternalism is 'inevitable' (2003: 1171) and that anti-paternalistic arguments are a 'nonstarter' (p.1165). People are indeed highly sensitive to the choice architecture, and to the extent that there does not exist a 'neutral' frame, any framing of a choice situation may influence the choice of the agent. It then falls to the person in charge of the

[7] If not, then it means that *x* is preferred to *y* if and only if the welfare associated to *x* is higher than the welfare associated to *y*. This representation of preferences by a real-valued function is possible only if preferences are complete and integrated.

[8] As a matter of terminology, the same idea has also been referred to as 'laundered preferences' or 'normative preferences', as well as 'experienced utility' in Kahneman *et al* (1997) – see Lecouteux and Mitrouchev (2021) for a detailed analysis of the definition of one's normative preferences.

design of the choice environment – the choice architect – to create the architecture of choice to improve the *navigability* of the environment (Sunstein 2019), and to help people to achieve what they truly want. Given that people follow possibly suboptimal heuristics, choice architects must design the choice environment so that following those heuristics ensures that the agent chooses *in fine* what maximises her welfare. Following Salvat (2014) and Whitman and Rizzo (2015), I will use the term 'behavioural paternalism' [BP] to refer to the various forms of paternalism that have been advocated based on the empirical findings of BE.[9]

The contractarian solution to the reconciliation problem

Unlike the welfarist narrative above, a *contractarian*[10] perspective on the reconciliation problem will not necessarily considers that BE challenges the sovereignty of agents to choose as they prefer. The normative justification of consumer sovereignty does not lie in the nature of the ends pursued by the agent, but merely in her freedom to choose as she desires. Robinson summarises this interpretation as follows:

> 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]: 50)

Whether individual preferences are integrated or not is not an issue for us, theoreticians: social institutions are deemed to be desirable to the extent that they allow agents to satisfy any preferences that *they might have*.

Sugden (2004, 2018) proposes within this perspective an *opportunity criterion*, according to which it is the opportunity sets of the agents which determine whether a situation is collectively preferred or not (in the sense of a social contract for which it is in the interest of each

[9] Note that BP does not overlap with BWE: Harrison and Ross (2018) 'quantitative intentional stance' is part of BWE, though they do not advocate paternalistic regulations. Furthermore, some non-welfarist approaches might also be considered to some extent as paternalistic, such as the boost program advanced by Grüne-Yanoff and Hertwig (2016, 2017) – even though the objectives differ from the traditional paternalistic arguments of the nudge agenda.

[10] I use the term 'contractarian' here because the argument of this section is mostly associated to Sugden's position. I offer a more detailed typology in Lecouteux (2021a).

citizen to accept it) – and not the satisfaction of some counterfactual preferences. An interesting property of the opportunity criterion is that it allows a reformulation of the first fundamental theorem of welfare economics, without requiring preferences to be integrated. The crux of the 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 integrated, 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).

While the welfarist justification considers that PS is a valid normative criterion under the condition that the agent's preferences are consistent with the formal requirements implied by complete and integrated preferences, the contractarian justification maintains that we should accept the multiplicity of criteria and views of what is a good life. The only cases which might be problematic are cases of *self-acknowledged* failures of self-control (e.g. heroin addicts), although Sugden considers that such cases of genuine problems of self-control are quite rare (Sugden 2017). From a contractarian perspective, the main normative question is not 'how should I live?' but 'how do we live together?', given the diversity of our conceptions of what matters. While the former question is ethical – and should be left to ethicists – the latter is political, for which economists can legitimately have a say.

Welfare and the autonomous consumer

The argument I will develop now is that the different justifications of PS discussed above implicitly advanced the idea that preference satisfaction is valuable under the condition that the agent is 'autonomous' (with however radically different notions of autonomy). I therefore first propose a definition of autonomy in terms of preference formation, and then highlight the different interpretations of the autonomous agent implicitly advanced in NBE, as the unitary self, the psychodynamic self, and the socio-historical self. As a guiding illustration, I will consider the case of Octave, whose excess weight is the direct consequence of his eating habits. Food behaviours are indeed now a central theme in discussions around nudging – the opening

illustration of *Nudge* is about the optimal location of food items in a cafeteria – though richer discussions about the *factors* leading to such behaviours are often neglected. The notion of autonomy I propose will precisely focus on this stage of preference formation.

Welfare and autonomy

'Welfare' is a central notion in normative economics, while – as discussed above – its definition is often neglected or confusing. It remains probably still less imprecise and elusive than the notion of 'autonomy', which is one of the many concepts in philosophy which lacks a consensual definition. My aim is here to provide a definition of welfare and autonomy that fits in the common language of economists in terms of preferences. I will for this purpose start from Sunstein's (1991) definition in terms of preference satisfaction and preference formation.

A normative issue raised by BE is that preferences appear to be *endogenous* to the choice problem, with e.g. the influence of the choice architecture on individual choices. This may question the normative relevance of PS since the *process* of how preferences are formed can be impacted by third parties or elements that the agent would consider as alien. As an extreme example, take Badger *et al* (2007) study of the willingness to pay a second dose of buprenorphine for long-time heroin addicts:[11] one of their central findings is that subjects were willing to pay a much higher price a second dose of BUP – whether the dose was supposed to be available the same day or five days later – when they were currently deprived (before getting their first dose) than when they were satiated (right after getting their first dose). In this kind of situation, the context clearly influenced the preferences of the subjects when they were asked to report their preferences for a second dose: we can therefore wonder whether PS constitutes an adequate criterion, knowing that the preferences were formed in a seemingly heteronomous way. As noted by Sunstein:

> It is notable that the great expositors of liberalism in the nineteenth and twentieth century are emphatic in the rejection of the view that satisfaction of existing preferences is adequate for purposes of ethics or politics. […] Mill's rejection of that view is especially

[11] Buprenorphine (BUP) is a commonly used substitute of heroin, and the focus on a *second* dose aims at measuring the willingness to pay a 'non necessary' dose. I indicate here their main findings, though it should be kept in mind that their study was only based on 13 subjects – with therefore a very limited statistical power.

> emphatic in his essay on Bentham, where he criticises Bentham for the view that "[t]o say either than man should, or that he should not, take pleasure in one thing, displeasure in another, appeared to him as much as an act of despotism in the moralist as in the political ruler." Mill, by contrast, emphasised the need to explore the influences "on the regulation of … affections and desires," and pointed to "the deficiencies of a system of ethics which does not pretend to aid individuals in the formation of their own character" [Mill, 1950, pp.68,70,71]' (Sunstein 1991: 6)

We can distinguish between two distinct stages which are of normative relevance: the process of *preference formation*, followed by the stage of *preference satisfaction*. The former corresponds to the set of factors and dynamics that give rise to the preferences we have (why Octave has a strong preference for sweet foods), and the latter to our actual choice once our preferences have been determined (which food Octave actually buys and eats). While consumer sovereignty has been understood as a question of preference satisfaction – let people choose as they prefer – it is less clear whether the agent should also be 'sovereign' over the stage of preference formation. This is precisely this notion of sovereignty over the processes of preference formation that I will designate by the term 'autonomy' (which cannot be reduced to a notion of negative freedom, as in Conly's *Against Autonomy* (2013)). Simply put, I propose that the level of 'welfare' is determined at the preference satisfaction stage, depending on the choice of the agent, and the degree of 'autonomy' is determined at the preference formation stage, depending on the degree of 'control' of the agent over the process of preference formation. The criterion of PS is then considered as normatively relevant only if preferences are formed autonomously – or conversely, PS must be rejected if the preferences are significantly formed by processes which are *alien* to the agent. Determining the boundary between what is acceptable and not thus depends on the ontological nature of the agent. I now compare three broad ontological views of the agent used in NBE and highlight that those diverging views imply different conditions to qualify as 'autonomous.'

The first view is common to most works in BWE and is more explicitly stated by Sunstein (1991) and Hausman (2012). It corresponds to the traditional view of 'the' self, as a distinct and centralised psychological or philosophical entity, which perdures throughout our lives. It is the true, authentic self, an 'inner citadel' (Christman 1989). It is seen as the seat of

rationality, and autonomy is traced to reasoning. Following the terminology proposed by Meyers (2005), I will designate this view of the self as the *unitary self*:

> The *unitary self* is the independent, self-monitoring, self-controlling self that has been pivotal to autonomy theory. As the seat of rationality and thus rational deliberation and choice, the self-as-unitary is often viewed as the ground for free will and responsibility (Meyers 2005: 29)

This 'unity' derives from the internal consistency of reasoning, based on logic, which constitute the hallmark of autonomy. The self-as-unitary view has however been largely criticised in philosophy for being (i) overly rationalistic and (ii) individualistic, and is challenged by 'post-modern' views of the self (e.g. Christman 2009: 48-56). I suggest that the two alternative conceptions which are used in NBE covers those two neglected facets of selfhood.

First, the self-as-unitary view overemphasises the place of rationality and critical self-reflection while disregarding the fundamental place of psychological mechanisms in our lives. Rather than defining ourselves as rational agents impaired by psychological biases, we humans are psychological beings, who may benefit from an additional ability of self-reflection. Unlike the Kantian approach of the self-as-unitary, the autonomous agent here depends on an integrated personality which does not need to derive from reason. This constitutes the view of the *psychodynamic self*, which is endorsed mostly by Sugden (2018) – who acknowledges his distinctive Humean view of psychology – as well as (to some extent) Dold and Schubert (2018), who emphasise the role of creativity in the self-definition of the individual. Second, the self-as-unitary view neglects the sociality of the individual, and that being embedded in a specific socio-historical environment fundamentally shapes one's preferences. The view of the *socio-historical* self is historically one of the earliest critiques of the PS criterion, as it lies at the core of Galbraith's critique of consumer sovereignty (Chirat 2020). This is the view advanced by Harrison and Ross (2018) and the philosophical framework of the 'quantitative intentional stance'.

Behavioural paternalism and the unitary self

The first strand of literature in NBE endorses the view of the unitary self, as in Sunstein's definition of autonomy:

> 'The notion of autonomy should refer … to decisions reached with a full and vivid awareness of available opportunities, with reference to all relevant information, and without illegitimate or excessive constraints on the process of preference formation' (Sunstein 1991: 11)

We can note here the similarity with Sunstein and Thaler (2003) later characterisation of counterfactually relevant choices (from a normative perspective): decisions that would have been made 'if [the agent] had complete information, unlimited cognitive abilities, and no lack of self-control' (Sunstein and Thaler 2003: 1162). Normative arguments based on the premise that BE reveals that real individuals make logical errors, or that being inconsistent is a genuine problem, implicitly rely on the self-as-unitary view. Such arguments indeed consider that PS can be applied if and only if the agent is autonomous in the sense of being logically consistent. This coincides with the model of the *inner rational agent* highlighted by Infante *et al* (2016).

Even though they do not call for a paternalistic agenda, this is also the implicit view of Bernheim and Rangel (2009) and Bernheim (2016), who reduce welfare analysis to choice data for which there is no ambiguity (i.e. no inconsistency). Bernheim (2016, p.49) identifies the decisions which are not welfare-relevant by examining 'evidence concerning the processes of observation, attention, memory, forecasting, and/or learning, with the object of determining the contexts in which certain types of facts are systematically ignored or processed incorrectly', i.e. evidence that the reasoning of the agent was heteronomous. Hausman and Welch (2010) also advance a similar definition of the autonomous self:

> The reason why nudges such as setting defaults seem … to be paternalist, is that in addition to or apart from rational persuasion, they may 'push' individuals to make one choice rather than another … [W]hen this 'pushing' does not take the form of rational persuasion, their autonomy – the extent to which they have control over their own

> evaluations and deliberations – is diminished. Their actions reflect the tactics of the choice architect rather than exclusively their own evaluation of alternatives

The autonomy of the agent is measured by its ability to critical self-reflection, and only rational persuasion is seen as an acceptable motive in the process of preference formation. This is in line with Hausman (2012) general view that preferences are the product of *reasoning*: even though he claims he does not have any substantive theory of well-being, Hausman nonetheless considers that preferences are 'more like judgments than feelings' (p.135), suggesting that preferences ought to be complete, transitive, and context-independent, as a matter of logical necessity. Being autonomous in this context is highly demanding and requires a high capacity of rational self-reflection – and BE offers direct evidence that human subjects probably lack the capacity for such autonomous thinking. It is therefore not surprising that the authors endorsing the self-as-unitary view are more willing to reject PS, since it is very likely that individual preferences were not formed in an autonomous way.

Consider Octave. His current food preferences are normatively relevant if and only if they are in line with the desires of his true self. It is supposed moreover that such desires must be the outcome of a rational deliberation (BP does not consider the possibility that one's true self could have volatile desires). Octave is allowed to have an unhealthy lifestyle, though only if it is justified by e.g. time preferences such that he exponentially discounts the future with a high discount rate. The long-term consequences of his behaviour would indeed be negligible compared to the immediate enjoyment of consumption. If, however, Octave's food preferences are determined by e.g. choice architecture (he does not really care about what to order for dessert and merely picks the first item), are the outcome of a present bias (generating time-inconsistent preferences), and more generally cannot be the outcome of a rational and context-independent deliberation, then Octave lacked autonomy and is likely to choose against his own interest. The typical interpretation of Octave's situation is that, deep inside, he wants to lose weight, but is unable to do so and to commit to a diet because of e.g. problems of self-control. The role of the benevolent choice architect is to improve the choice environment so that it becomes easier for Octave to make the 'right' choices – limit calorie-intake, encourage activities burning calories, etc.

Note that the justification of BP has a very peculiar interpretation of BE findings. It is postulated that the preferences of the individual, if critically assessed from the perspective of the autonomous self-as-unitary, should eventually conform to the requirements of rational choice theory. That is, when considering the possibly conflicting and incomplete preferences that the agent may have when first confronted to a choice problem, it is supposed that the autonomous agent has access to a certain algorithm allowing her to solve a complex multi-dimensional problem into a simple complete and integrated preference relation, her true preferences. Infante *et al* (2016) challenge this assumption, by stressing that 'we know of no argument, either in behavioural economics or in the theory of rational choice, that would justify the assumption that such an algorithm exists' (p.17). I have also argued elsewhere that this assumption lacks both empirical and normative justifications, and that BP fails to justify why the autonomous agent would abide by the standards of coherence of rational choice theory (Lecouteux 2021b). We arrive at the uncomfortable statement that one's subjective values *must be consistent* with the formal requirements of rational choice theory, which leaves little room for the 'as judged by themselves' clause advanced by Sunstein and Thaler.

Furthermore, the conditions required to qualify as autonomous are probably unreachable by any human being (think of the 'unlimited cognitive abilities'), so we cannot confidently state that such counterfactual reasoning – that nobody is able to run properly – would indeed lead to the formation of complete and integrated preferences (see Qizilbash 2012 for a similar argument). We therefore must postulate the outcome of such algorithm without being able to check its solution. Even though everyone is very likely to prefer living a long life in good health rather than a shorter one, how can we be *certain* that Octave's true self would necessarily – and unambiguously – prefers healthier habits? If Octave's true self deeply values spontaneity and the possibility to change one's mind, then even a logically coherent Octave could choose to opt for possibly incoherent habits.

The psychodynamic self

By opposition to the rationalistic approach of the unitary self, Sugden (2018) proposes a psychodynamic conception inspired from Hume's experimental psychology (see Sugden 2020). He rejects reason-based account of preferences and acknowledges their psychological nature.

Preferences are the outcome of our volatile desires, which are conscious or not. Meyer (2005) designates a related view as the *divided self*:

> Split between consciousness and self-awareness, on the one hand, and elusive unconscious affect and desire, on the other, the self-as-divided is characterized by inner depth, complexity, and enigma. The fluid but distinctive psychocorporeal economy of the self-as-divided is manifest in a unique – indeed, a vibrantly individualized – personality. In an important respect, the value we place on autonomy pays tribute to this conception of the self, for autonomy enables people to express their individuality in the way they choose to live.

Unlike within the self-as-unitary view, psychology is not interpreted as interfering with a latent rationality. Rationality is rather an *add-on* to human psychology – this is incidentally the type of evolutionary arguments that justify dual systems of cognition. Kahneman (2011: 24) for instance suggests that '[when] System 1 runs into difficulty, it calls on System 2 to support more detailed and specific processing that may solve the problem of the moment.' While such models are often used in BWE to justify the contrast between the 'rational' and 'psychological' parts of our agency, it must be kept in mind that the deliberative/rational system is here to *supplement* the automatic/psychological system, which constitutes the core of the individual's identity. Humans are better described as psychological beings with an ability of self-reflection, rather than as 'faulty Econs', i.e. rational beings impaired by psychological biases.

Within this view, threats to autonomy are much weaker than considered in BP. The autonomous agent is united through its distinctive personality, as a 'self-acknowledged locus of responsibility' (Sugden 2004: 1018), and must take responsibility for his past, present and future actions without necessarily referring to a benchmark in terms of internal consistency. There could however still be cases posing serious problems in the process of preference formation, mostly if changes in preferences are driven by largely unconscious – and therefore unnoticed – factors, for which the agent cannot take responsibility. This is the case of pathological disorders and situations of addiction, for which there may exist an explicit conflict between the passions riding one's behaviour and the passive observation of this behaviour by one's reason. Apart from situations of eating disorder, there is probably no problem with Octave food behaviour, even though he may experience the perfectly normal coexistence of conflicting preferences for sweets

and staying fit. As long as he recognises those desires as his own and part of his personality, being incoherent by neoclassical standards does not pose any normative challenge.

Apart from cases of fraud or explicit manipulations by third parties (which may constitute a violation of the rules of fair competition in an economic setting), or pathological conflicts within the individual, the psychodynamic self – as a self-acknowledged person with its distinctive personality – is often autonomous with respect to the process of preference formation. This is the reason why such views will not see in BE a fundamental normative issue.

The socio-historical self

A third view focuses on the inherent sociality of selves: the individual is embedded in a specific socio-historical environment, and it might not make sense to debate the nature of preferences while extracting the agent from this context. The *socio-historical* view considers the socialised or enculturated self, shaped by culturally transmitted norms and values. It is the conception underlying Harrison and Ross (2018) Dennettian framework, the quantitative intentional stance towards BWE. 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: 20)

Dennett advances a definition of the self in terms of narrativity, according to which the individual must take the intentional stance toward herself, to 'try to make all of [her] material cohere into a single good story', the person's 'autobiography' (Dennett 1992: 114). Preferences and beliefs are more than a useful tool for the theoretician to represent peoples' behaviours, they constitute our *shared language* required in the process of socialisation. When attributing a utility function and subjective beliefs through standard Bayesian techniques, we are describing the ecology within which the individual evolves rather than some inner mental states. Normative judgements about individual preferences are judgements about the dynamics of the system that contributes to shape those preferences. If Octave is known as a 'bon vivant' – and that he himself

contributes to maintain this identity through his behaviour – then his food behaviour should be seen as the consequence of a social system that may value and sustain this type of lifestyle.

If the individual is the product of a complex environment, it becomes challenging to define a satisfactorily notion of 'autonomy.' I will here endorse Christman (2009) definition in terms of competence and authenticity:

> Autonomy involves competence and authenticity; authenticity involves non-alienation upon (historically sensitive, adequate) self-reflection, given one's diachronic practical identity and one's position in the world. (Christman 2009: 155)

A simple interpretation is that being autonomous means being *aware* of the main factors driving one's behaviours, and also *accepting* those factors. Behavioural interventions such as nudges, when they remain unnoticed, are therefore a threat to the individual's autonomy, because they induce behavioural changes that the agent is not aware of, leading to choices for which she cannot take responsibility. This definition of individual autonomy is a *responsiveness-to-reasons* account of autonomous agency (e.g. Wolf 1990, Fischer and Ravizza 1998), according to which 'an agent does not really govern herself unless her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does' (Buss 2014). This is distinct from the unitary self, which is associated to a *responsiveness-to-reasoning* account of autonomous agency (Christman 1991). While the latter requires the agent to critically evaluate her motives and reasons on the basis of her beliefs and desires, and to adjust them in the light of her evaluation and the principles of logic (which should remove any logical inconsistency), the former does not constraint how the agent should process new reasons in her evaluation. There is a distinction between improving the *instrumental process* through which the agent arrives at a decision, and the *inputs of the process*, without constraining how the agent might arrive at her decision.[12]

[12] In both cases, we face the difficulty of defining the counterfactual outcome of a process of self-reflection, on which an external observer has only an imperfect knowledge. It is noteworthy nevertheless that although Christman's notion of autonomy relies on an ability of self-reflection, it is still far from the view of the unitary self. The agent we are considering (whose autonomy can be bounded) is not a counterfactual agent defined as what the individual would be if she were ideally rational: the agent is the actual individual, aware that her

Consider now Octavia, who – unlike Octave – strictly follows a healthy diet and never deviates from it, because she cares a lot about her fitness. From the perspective of BP, Octavia is perfectly autonomous, and we should be happy with her preferences. However, if Octavia behaves as she does because she lives in a society with a strong culture of fatphobia, specifically towards women, and that being a 'bon vivant' is only a positive feature for men, then it is less clear that Octavia's preferences have been formed autonomously. The criterion is here that, if Octavia was aware of such influence, then she would find it acceptable upon reflection. It is far from certain that it would be the case, which could give a new justification for public policies. The objective would not be to help Octavia to satisfy preferences she does not have (by incentivising her to deviate from her diet), but rather to alter the social institutions to make sure that situations like a norm of fat shaming are less likely. Promoting autonomy at the stage of preference formation could thus justify more ambitious policies than nudging, though such policies would be of a very different nature.

From the ontology of the self to public policy

The most vibrant advocates of BP are the authors endorsing the unitary view of the self, which is however limited since it fails to capture both the inherent psychology and sociality defining the agent. The psychodynamic conception generally does not question the legitimacy of our possibly conflicting desires – apart from explicit cases of manipulation and pathological disorders – and therefore maintain the normative relevance of the consumer sovereignty principle. The socio-historical view however, while acknowledging that autonomous agents can form preferences that deviate from rational choice theory – as long as they are aware of the causes and accept them – may also offer an alternative justification for regulations. Indeed, the fact that people's preferences could be manipulated by third parties, for their personal, commercial or political

---

behaviour is driven by many psychological processes, on which she has a limited control, but who nevertheless accepts to live with this fact of human psychology. We do not need to refer to a hypothetical agent with unlimited cognitive abilities to test whether the autonomy of the individual is limited, but could simply rely on the judgement of real individuals to have an idea of this counterfactual outcome.

interests – unlike the benevolent nudgers and choice architects central to BP – constitutes a serious threat to their autonomy.

The argument that PS may not constitute an adequate normative criterion because of the endogeneity of preferences was one of the earliest critiques of the consumer sovereignty principle formulated by Galbraith (1938). Similarly to current debates, Galbraith rejected the idea of a difference between 'rational' and 'irrational' preferences put forward by Kahn (1935) – which is the core argument in today's BP. 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: 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. What behavioural economics tells us is not necessarily that our preferences are only marginally impacted by the context – as in standard arguments discussing the design of the choice architecture – but are more fundamentally *shaped* by our social environment. This idea has been developed in cognitive sciences through the *mindshaping* hypothesis (Zadwiski 2013), according to which it is our collective ability to *shape* each other's minds that is the key to our evolutionary success. It is defined by Zawidski (2013: xiii) as 'a group accomplishment, involving simultaneously interpretive and regulative frameworks that function to shape minds.' It includes social devices such as norms, conventions, or institutions, which regulate individual behaviours and social practices. The aim of such devices is to generate a 'cognitive homogenisation' (Zawidski, 2013: 65), so that people's behaviours become easily interpretable. Preferences are thus an evolving social product of the economic system: when studying revealed preferences (e.g. the preference for certain goods, equilibria, institutional arrangements) we are primarily looking at the economic system rather than the agents who are part of it.

As an extension of Octave's situation discussed throughout the chapter, consider more generally the question of obesity. The main narrative of BP is that obese people have a problem of self-control, and that they fail to commit to their diet. This discourse tends to pathologize the agent, making her (and her deficient psychology) responsible for her current situation. Obesity can however also be analysed as the outcome of poor socio-economic conditions (Drewnowski & Specter 2004, Tanumihardjo *et al* 2007), as well as misaligned economic incentives (see Galizzi 2012: 19 on the relatively low price of calories in junk food). Furthermore, one of the

factors of the current obesity pandemic has been how the food industry managed to convince consumers of the higher dietary quality of 'low-fat' products, which were flavoured with sugar and artificial sweeteners (Nestle, 2013). An effective strategy against obesity would therefore require altering *social institutions* – which may later impact the process of preference formation – and not merely focusing exclusively on the demand side of the market.

I have argued in this chapter that current disagreements about the problem of reconciling normative and behavioural economics probably stem from different ontological conceptions of the agent. Most of BP keep an atomistic view of the agent and consider that PS offers a satisfactory normative criterion only if the agent is autonomous in the sense of the autonomous self-as-unitary. This position is at odds with the more psychological view of the agent in the psychodynamic interpretation – according to which BE does not pose a radical challenge to the consumer sovereignty principle – as well as with the socio-historical view that emphasises the need to consider the role of institutions in the process of preference formation. These divergences highlight that we should not expect a simple 'solution' to the reconciliation problem, simply because we cannot expect behavioural economists to agree on a single conception of the self. Depending on the context, one facet of the self might be more relevant (e.g. the unitary self if I have to choose between identical options with complex pricing systems), which will condition the criteria to test whether the agent is autonomous, and the types of policy which could be adequate, if any. Even though there may be disagreements about the relevant ontological conception of the self, I would like to conclude by stressing the need to investigate further the dynamics of preference formation. Indeed, as soon as preferences are endogenous – and this is the core finding of BE – normative claims about PS cannot be avoided. In the words of Scitovsky:

> And here I should like to end on a note I started out with: the economist could wash his hands of value judgments only if the public's preferences were really given and he could accept them as such. As soon as this ceases to be true and the public's preferences are influenced by economic agents and the economic environment, value judgments on whether this influence is good or bad, in need of restraint or reform, cannot be avoided. If the economist feels incompetent to make such judgments himself, he should at least

admit their legitimacy and provide the analytical framework to help others to make these judgments (Scitovsky, 1962: 268)

References

Badger, G. J., Bickel, W. K., Giordano, L. A., Jacobs, E. A., Loewenstein, G., & Marsch, L. (2007). Altered states: The impact of immediate craving on the valuation of current and future opioids. *Journal of health economics*, *26*(5), 865-876.

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

Buss, S. (2014). "Personal Autonomy". In Zalta, E. N., editor, The Stanford Encyclopedia of Philosophy. Winter 2014 edition

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.

Christman, J. (1989). The inner citadel: Essays on individual autonomy.

Christman, J. (1991). "Autonomy and Personal History". Canadian Journal of Philosophy, 21:1–24.

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

Conly, S. (2013). Against Autonomy. Justifying Coercive Paternalism. Cambridge: Cambridge University Press.

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.

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

Drewnowski, A., & Specter, S. E. (2004). Poverty and obesity: the role of energy density and energy costs. *The American journal of clinical nutrition*, *79*(1), 6-16.

Fischer, J. and Ravizza, M. (1998). Responsibility and Control: A Theory of Moral Responsibility. Cambridge: Cambridge University Press.

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

Galizzi, M. M. (2012). Label, nudge or tax? A review of health policies for risky behaviours. *Journal of public health research*, *1*(1), 14-21.

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

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

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

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. and Welch, B. (2010). "Debate: To Nudge or Not To Nudge". Journal of Political Philosophy, 18:123–136.

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

Heukelom, F. (2014). *Behavioral economics: A history*. Cambridge University Press.

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

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

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.

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

Lecouteux, G. (2021a). 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*, 208-233.

Lecouteux, G. (2021b). Behavioural welfare economics and consumer sovereignty, In Heilman, C. & Reiss, J. (Eds), *Routledge Handbook of Philosophy of Economics*, 56-66.

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

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

Meyers, D. T. (2005). Decentralizing autonomy: Five faces of selfhood. *Autonomy and the Challenges to Liberalism*, 27-55.

Nestle, M. (2013). *Food politics: How the food industry influences nutrition and health* (Vol. 3). Univ of California Press.

Pareto, V. (1909). Manual of Political Economy. London: McMillan. Translated by A. Schwier from the 1927 french edition (1971).

Qizilbash, M. (2012). Informed desire and the ambitions of libertarian paternalism. *Social Choice and Welfare*, *38*(4), 647-658.

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

Salvat, C. (2014). Behavioral paternalism. *Revue de philosophie economique*, *15*(2), 109-130.

Scitovsky, T. (1962). On the principle of consumers' sovereignty. *The American Economic Review*, *52*(2), 262-268.

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

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.

Sugden, R. (2020). Hume's experimental psychology and the idea of erroneous preferences. *Journal of Economic Behavior & Organization*.

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

Sunstein, C. R. (2019). *On freedom*. Princeton University Press

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.

Thoma, J. (2021). In defence of revealed preference theory. *Economics & Philosophy*, *37*(2), 163-187.

Tanumihardjo, S. A., Anderson, C., Kaufer-Horwitz, M., Bode, L., Emenaker, N. J., Haqq, A. M., ... & Stadler, D. D. (2007). Poverty, obesity, and malnutrition: an international perspective recognizing the paradox. *Journal of the American Dietetic Association*, *107*(11), 1966-1972.

Whitman, D. G., & Rizzo, M. J. (2015). The problematic welfare standards of behavioral paternalism. *Review of Philosophy and Psychology*, *6*(3), 409-425.

Wolf, S. (1990). Freedom within Reason. New York: Oxford University Press

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