Routledge
Taylor & Francis Group

Journal of Economic Methodology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjec20

Permissible preference purification: on context-dependent choices and decisive welfare judgements in behavioural welfare economics

Måns Abrahamson

To cite this article: Måns Abrahamson (15 Sep 2023): Permissible preference purification: on context-dependent choices and decisive welfare judgements in behavioural welfare economics, Journal of Economic Methodology, DOI: 10.1080/1350178X.2023.2257212

To link to this article: https://doi.org/10.1080/1350178X.2023.2257212

Published online: 15 Sep 2023.

JOURNAL OF ECONOMIC METHODOLOGY
https://doi.org/10.1080/1350178X.2023.2257212

Routledge
Taylor & Francis Group

Permissible preference purification: on context-dependent choices and decisive welfare judgements in behavioural welfare economics

Måns Abrahamson

Erasmus School of Philosophy, Erasmus University Rotterdam, Rotterdam, The Netherlands

ABSTRACT

Behavioural welfare economics has lately been challenged on account of its use of the satisfaction of true preferences as a normative criterion. The critique contests what is taken to be an implicit assumption in the literature, namely that true preferences are context-independent. This assumption is considered not only unjustified in the behavioural welfare economics literature but unjustifiable – true preferences are argued to be, at least sometimes, context-dependent. This article explores the implications of this 'critique of the inner rational agent'. I argue that the critique does not support a wholesale shift away from the use of true preferences as an evaluative standard in normative economics; instead, the critique implies that behavioural welfare economists need to inquire into and establish the 'source' of particular context-dependent choices in individuals' decision-making. The source determines the permissibility of correcting individuals' context-dependent choices and can, in some situations, support decisive welfare judgements.

ARTICLE HISTORY
Received 16 November 2022
Accepted 6 September 2023

KEYWORDS
Behavioural welfare economics; preference purification; context-dependence; paternalism; nudge

1. Introduction

Behavioural welfare economics is an approach to welfare assessment and economic policymaking that underlies much of the recent work in normative economics. The approach seeks to reconcile (i) behavioural economic evidence that individuals many times behave inconsistently with the assumption of maximising a stable and context-independent utility function (see, e.g. Camerer & Loewenstein, 2004; DellaVigna, 2009), with (ii) a subjectivist, choice-based normative framework.

This is done by way of 'preference purification' (Sugden, 2018b). Rather than basing welfare judgements on individuals' revealed preferences, which are often context-dependent and therefore lead to ambiguous welfare judgements (Hausman, 2022b; Hédoin, 2017), behavioural welfare economics seeks to simulate and satisfy preferences that individuals *would* act upon under ideal reasoning circumstances, where they pay 'full attention and [possess] complete information, unlimited cognitive abilities, and complete self-control' (Thaler & Sunstein, 2008, p. 5). In other words, the core idea of behavioural welfare economics is to reconstruct the preferences individuals would reveal had they not been subject to reasoning errors and to use these so-called true preferences as a basis for welfare judgements. This purification of preferences can be approached by excluding preferences revealed in contexts where the conditions of ideal reasoning are not satisfied from the set of welfare-relevant choices in favour of preferences that are (e.g. Bernheim, 2016), or by recreating welfare-relevant preferences by removing the effects of different biases or reasoning

CONTACT Måns Abrahamson abrahamson@esphil.eur.nl

imperfections from individuals' preferences after they have been elicited (e.g. Lipman et al., 2019; Pinto-Prades & Abellan-Perpiñan, 2012).

Individuals' true preferences are assumed to have authority on matters pertaining to their well-being and are invoked to make decisive welfare judgements in situations of observed context-dependent choice. Importantly, true preferences are understood as subjective in nature – they are taken to capture what makes individuals 'better off, *as judged by themselves*', rather than third-party welfare judgements (Thaler & Sunstein, 2008, p. 5). Therefore, in purifying individuals' preferences to support decisive welfare judgements, the assumption in behavioural welfare economics is that targeted individuals' subjective interests are respected.

The coherence of this approach to normative economics has recently been challenged. The critics argue that behavioural welfare economics relies on an implicit assumption that true preferences are context-independent, based on a psychologically and philosophically problematic as-if model of the human being as having 'a neoclassically rational inner agent, trapped inside and constrained by an outer psychological shell' (Infante et al., 2016b, p. 22), or similarly that 'each person has a neoclassical agent deep within themselves struggling to surface' (Whitman & Rizzo, 2015, p. 423). Behavioural welfare economics is therefore taken to be fundamentally flawed: the use of true preferences to arbitrate between individuals' context-dependent choices imposes unfounded assumptions about latent rationality on targeted individuals, with the upshot that associated welfare judgements do not respect those individuals' own judgements about well-being; this, the argument goes, amounts to an objectionable form of paternalism.

This article examines the implications of this 'critique of the inner rational agent' for behavioural welfare economics. I argue that a categorical dismissal of behavioural welfare economics based on the critique is unwarranted – the critique only establishes that true preferences will not *necessarily* be context-independent and therefore that the context-independence of true preferences should not be taken for granted. The upshot is that behavioural welfare economic approaches that make a blanket assumption of true preferences being context-independent are problematic in the ways specified by the critique; however, approaches that instead treat context-independence as a contingent feature of true preferences are not subject to the critique – they do not presuppose a problematic 'inner rational agent'.

In light of this, I argue that the central implication of the critique is that behavioural welfare economists need to inquire deeper into, and distinguish between, different 'sources' of observed context-dependent choice in individuals' decision-making. Some context-dependent choices can be attributed to variation in individuals' evaluative perspective between contexts, while other context-dependent choices can be attributed to false beliefs or failures of self-control caused by contextual cues. I argue that context-dependence due to false beliefs or self-control failure can permissibly be purified in light of individuals' subjective interests, thereby supporting decisive subjectivist welfare judgements in associated situations of context-dependent choice.

While the distinction between different sources of context-dependent choice allows for permissible preference purification in *principle*, epistemological concerns remain. I argue that empirical procedures are available to behavioural welfare economists to discern sources of context-dependence, which supports permissible preference purification in *practice*.

The aim of this article is then to show that behavioural welfare economics has access to viable conceptual and methodological resources to support permissible purification of individuals' preferences, and with that decisive welfare judgements, in a number of situations of context-dependent choice. This has implications for associated policy programmes, such as 'nudging' (Thaler & Sunstein, 2008) and 'boosting' (Hertwig & Grüne-Yanoff, 2017), that draw on ideas of true preferences to warrant paternalistic policymaking. It is, however, important to note that there is a relevant step between making decisive welfare judgements based on true preferences and intervening in individuals' choices in light of these welfare judgements. This article provides a partial justification, based on subjective welfare considerations, for steering some of people's choices by way of policy intervention. A complete justification for policy intervention should consider additional concerns, such as

whether a particular intervention would undermine individuals' autonomy and dignity, or would be manipulative in problematic ways (see Schmidt & Engelen, 2020 for an overview on the ethics of nudging).

The article proceeds as follows. Section 2 explicates the critique of the inner rational agent. Section 3 discusses the implications of the critique. Section 4 delineates three broad sources of context-dependence, highlights important psychological mechanisms and cases of context-dependent choice in relation to these sources, and discusses the permissibility of correcting associated context-dependence. Section 5 elaborates upon a problem of underdetermination when distinguishing between sources of context-dependence and argues for the availability of empirical procedures that can support credible attributions to sources in practice. Section 6 concludes.

2. The critique of the inner rational agent

The core of the critique of the inner rational agent is that behavioural welfare economics relies on an unjustified, and indeed unjustifiable, implicit assumption that individuals' true preferences – the preferences individuals *would* reveal under ideal reasoning circumstances – are context-independent. The critique was forcefully introduced in Infante et al. (2016b), but variations or complementary versions of the critique can also be found in Whitman and Rizzo (2015), Rizzo and Whitman (2020), Infante et al. (2016a), Sugden (2015, 2017, 2018a, 2018b), and Dold (2018).

Implicated in the critique is a large number of normative works in economics making use of behavioural economic insights. With significant overlap between the critics, explicitly targeted works include, for example: libertarian paternalism, as introduced in Sunstein and Thaler (2003) and popularised in Thaler and Sunstein (2008), which seeks to improve individuals' well-being through freedom-preserving policy interventions that harness cognitive biases and reasoning limitations; the 'unified framework' developed in Bernheim and Rangel (2007, 2009) and reinterpreted in Bernheim (2016), which seeks to extend the standard choice-theoretic framework of welfare analysis to situations of context-dependent choice; Köszegi and Rabin's (2007, 2008) 'revealed mistakes' approach, which seeks to identify systematic mistakes in individuals' choices in order to reconstruct their welfare-relevant preferences; and preference purification, as put forward by Hausman (2012), which seeks to extend the domain where individuals' preferences can be considered reliable indicators of well-being by correcting for mistakes in their reasoning.

What these, and other, behavioural welfare economic approaches have in common is that they seek to make decisive welfare judgments in situations of observed context-dependent choice. These are situations where individuals' choices are affected by features of the decision setting that 'have little or no relevance to [their] well-being, interests, or goals' (Infante et al., 2016b, p. 2), and that should therefore be 'irrelevant to the decision' (p. 18). This includes 'framing effects', where redescriptions of available options, conveying the same information, affect choice (e.g. Tversky & Kahneman, 1981); some forms of 'menu dependence', where the inclusion or exclusion of additional options in the choice set affect choice (e.g. Simonson & Tversky, 1992); and broader 'environmental factors', such as whether, for instance, a particular background music is playing when making a choice (e.g. North et al., 1997).[1]

An oft-invoked instance of context-dependent choice is Sunstein and Thaler's (2003) example of the 'cafeteria'. In the cafeteria, customers have a tendency to choose food items that are presented earlier in the cafeteria line: if cakes are presented earlier in the line, customers tend to opt for cake; if apples are presented earlier in the line, customers tend to opt for apples. The arrangement of food items in the cafeteria should, plausibly, be irrelevant to individuals' decision of what item to consume. Yet it affects their choices.

In order to make decisive welfare judgements in situations of context-dependent choice (e.g. the cafeteria), and to justify related paternalistic policymaking such as nudging individuals in the direction of a particular option (e.g. apples), behavioural welfare economists appeal to individuals' true

preferences: 'What *would* individuals' preferences over the options be under ideal reasoning circumstances?'. This exercise can only support decisive welfare judgements if individuals' true preferences are context-independent. If, for example, individuals' true preferences were affected by the arrangement of food items in the cafeteria in the same way their revealed preferences are, the behavioural welfare economic approach would not get off the ground.

The critics argue that this assumption of true preferences being context-independent remains *unjustified*. In behavioural welfare economics, true preferences are assumed to be subjective. That is, true preferences, even though idealised, are assumed to reflect individuals' own judgements about their own good. Infante et al. (2016b) suggest that this entails that individuals have 'potential access to some mode of *latent reasoning* that generates subjective preferences that satisfy conventional principles of rational consistency' (p. 10). However, proponents of behavioural welfare economics 'do not explain what that mode of reasoning is or how it generates coherent preferences' (p. 10), which means that 'rational choice itself – represented by error-free reasoning of the inner rational agent – is not given any psychological explanation' (p. 15). In other words, the assumption of context-independent true preferences lacks a justification.

Additionally, the critics argue that the assumption of context-independent true preferences is, at least sometimes, *unjustifiable*. This is argued to be the case by demonstrating that individuals may choose in a context-dependent manner without violating any norm of sound reasoning, which implies that their true preferences would, in these instances, also be context-dependent.

The starting point for the critics' justifications for context-dependent true preferences is that individuals' preferences are often formed in the process of making a choice. Both Rizzo and Whitman (2020, p. 58) and Dold (2018, p. 165) cite Buchanan (1979/1999) in that '*not even* individuals have well-defined and well-articulated objectives that exist independently of choices themselves' (p. 258). That is, many times individuals do not hold pre-settled preferences over particular options that they appeal to, or discover, when making a decision – instead, individuals 'construct' their preferences on the fly (see Lichtenstein & Slovic, 2006).

Assuming that individuals' preferences are constructed on the basis of values that they hold, there are a number of reasons why this might lead to context-dependent choice without violating any norm of sound reasoning. First, individuals' choices may be context-dependent without violating any norm of sound reasoning when the relative importance individuals assign to their different values are partially determined by the context and available options promote a number of different values. For example, in the cafeteria, an agent's decision may be governed by the values 'being healthy' and 'enjoying pleasurable experiences'. The agent then forms an overall preference by weighing these different values against each other. As Infante et al. (2016a) note, it may be the case that 'this weighing operation is influenced by contextual cues', since 'the relative importance of different dimensions of [individuals'] lives depend on what they are currently attending to' (p. 36). If cake is placed prominently, the attribute of 'pleasurableness' is more salient and therefore given more weight in the balancing of values; if apples are placed more prominently, the attribute of 'healthiness' is more salient and therefore given more weight in the balancing of values. Nothing seems to have gone wrong in the agent's reasoning in this instance – the context-dependence is due to imprecision in the values that the agent uses as starting point for her reasoning. The implication being that the agent's true preferences would, in this instance, be context-dependent in the same way.

Second, individuals' choices may be context-dependent without violating any norm of sound reasoning when individuals are unable to trade off the different values that govern a specific decision and thereby to form an overall preference. For example, the agent may not, for whatever reason, be able to assign relative value to 'being healthy' and 'enjoying pleasurable experiences' in relation to the options in the cafeteria. In instances like this, when preferences are incomplete, the agent may make a decision on the basis of her inclinations, which may be affected by contextual cues – such as the salience of the options. Infante et al. (2016b) argue that this may also hold under ideal reasoning circumstances, since sound reasoning does not necessarily generate a complete

preference ranking (cf. Hausman, 2012, p. 19). If completeness is not an axiom, but a 'boundary condition' of sound reasoning, 'context-dependent choices are not necessarily mistakes that can be corrected by purification' (Infante et al., 2016b, p. 13).

In a similar vein, Rizzo and Whitman (2020) argue that it can be reasonable for individuals to make context-dependent choices due to 'preference rotation'. If a decision is governed by multiple values, and different options promote these values to varying degrees, choosing one option consistently would lead to some of the agent's values 'getting short shrift' (p. 60). For example, consistently choosing apples in the cafeteria would neglect the value 'enjoying pleasurable experiences'. A reasonable solution to this, Rizzo and Whitman argue, is for the agent to 'rotate' her preferences over the options – 'sometimes favouring one, sometimes favouring another' (p. 60) – in order to facilitate 'balance between underlying values' (p. 63). Since the success criteria for fulfilling particular values are often somewhat vague (e.g. 'healthy' is a vague predicate), when the agent ought to rotate her preferences is presumably not an exact science. This suggests that contextual cues (such as salience) can influence individuals' choices in these instances without there being a failure of reasoning.

Based on the considerations above,[2] the critics take themselves to have provided a fundamental 'methodological' (Infante et al., 2016b) or 'ontological/metaphysical' (Dold, 2018; Whitman & Rizzo, 2015) critique of behavioural welfare economics. But what does this critique imply for the viability of the behavioural welfare economic approach?

3. Implications of the critique of the inner rational agent

There is some ambiguity in the literature about what the critique of the inner rational agent implies for behavioural welfare economics. This section introduces and dismisses a possible reading of the implications of the critique before introducing a more plausible formulation of what conclusions to draw from the critique of the inner rational agent.

On one reading of the critics, the conclusion that can be seen to be drawn from the analyses is that behavioural welfare economics is fundamentally flawed and should be abandoned in favour of an alternative normative framework to welfare economics – for example, Sugden's 'opportunity criterion' (see, e.g. Sugden, 2004, 2018b), as suggested by Infante et al. (2016b), or a focus on the institutional environment in which individuals' preferences are formed (rather than on preferences themselves), as suggested by Dold (2018).

This reading is hinted at in several passages in the critical works. For example, after claiming that 'the idea that context-dependent choices are caused by errors of reasoning is fundamentally misconceived' (Infante et al., 2016b, p. 2), and that 'latent [i.e., true] preference is not a useful concept for normative economics' (p. 18), Infante and co-authors suggest that there is a need for 'a normative economics that does not presuppose a kind of rational human agency for which there is no known psychological foundation' (p. 22). This point is echoed by Dold (2018), who submits that his approach, in contrast to behavioural welfare economics, 'does not presuppose any kind of inner rational agency for which there is no convincing psychological or normative foundation' (p. 170). Whitman and Rizzo (2015) similarly suggest that there is a *non sequitur* at the heart of behavioural welfare economics of 'identifying an inconsistency, and then resolving it by designating one set of preferences as "true"' (p. 423), driven by a faulty metaphysical assumption that individuals (have the capacity to) hold context-independent true preferences.

This *categorical dismissal* of behavioural welfare economics, used to justify a wholesale shift away from using true preferences as normative criterion, is not supported by the critics' analyses. The critics establish that individuals may *sometimes* choose in a context-dependent manner without violating any norm of sound reasoning, which implies that individuals' true preferences may *sometimes* be context-dependent. This is, as was shown in the previous section, the case in multi-attribute decision problems when the relative weight individuals assign to their different values are partially determined by context, or when individuals are unable trade off their different values. But these

explanations do not exhaust the different reasons for why individuals make context-dependent choices. As I argue in the following section, false beliefs and self-control failure can also generate context-dependent choice and there are in these instances viable subjectivist grounds for purifying individuals' preferences.

Rather than a categorical dismissal of behavioural welfare economics, the critique is more plausibly understood as identifying two broad implications for the discipline. First, that the context-dependence of true preferences should not be taken for granted. As highlighted by the critics' analyses, true preferences will not *necessarily* be context-independent. Behavioural welfare economic approaches that assume that true preferences are always context-independent are problematic as specified by critics: they 'presuppose a kind of inner rational agency for which there is no convincing psychological or normative foundation'. An example of this is the behavioural welfare economic framework of Rubinstein and Salant (2012), which takes as its starting point the assumption that 'the *welfare* of an individual is reflected by an unobservable ordering (an asymmetric and transitive binary relation that relates every two alternatives)' that is systematically distorted by welfare-irrelevant details (p. 376).

However, behavioural welfare economic approaches do not have to presuppose that true preferences are always context-independent; instead, they can treat context-independence as a contingent feature of true preferences that may or may not hold in particular decision settings. Bernheim's (2016) 'unified framework' is an example of this. The framework operates by restricting the set of welfare-relevant choices on the basis of given criteria (in particular, removing choices based on mistaken characterisations of options) while leaving open whether this leads to an internally consistent set of choices. As noted by two recent defences of behavioural welfare economics (Bernheim, 2021; Thoma, 2021): such an approach is not subject to the critique as it does not presuppose 'an inner rational agent'.

A second implication of the critique, following the first, is that behavioural welfare economics needs to inquire deeper into, and distinguish between, different reasons for observed context-dependent choice. Behavioural welfare economists need to convincingly establish that observed context-dependence in *any particular setting* of interest – not in general – can be attributed to aspects of individuals' preference formation and implementation that would be *removed* in a process of permissible purification. I argue that this entails that behavioural welfare economists need to locate the 'source' of given context-dependent choices in individuals' decision-making – whether the context-dependence is due to variation in individuals' evaluations, false beliefs, or self-control failure. This is essential for the approach, as the source determines the *permissibility* of arbitrating between observed context-dependent choices.

'Permissible' preference purification is here understood as disregarding or correcting some of people's context-dependent choices while respecting and better promoting the targeted individuals' own, subjective interests. Two criteria can then be seen to govern the permissibility of preference purification. First, that any correction of context-dependence must be conducted in a way that respects targeted individuals' subjective interests. In line with a long tradition of anti-paternalism in economics, behavioural welfare economics seeks to avoid imposing external views on what constitutes a good life for people, and instead 'make choosers better off, as judged by themselves'. Adhering to this principle requires a clear conception of what constitutes these judgements – what individuals' subjective interests consist in. Throughout the rest of the article, individuals' subjective interests are identified with their values. More specifically, an agent's preferences (actual or hypothetical) are taken to reflect her subjective interests to the extent that they are grounded on evaluations of options' attributes, which the agent identifies with at the moment of choice, that are within the 'perimeters' of her values – a notion I borrow from Engelen and Nys (2020). The broad idea being that we all 'have values and things we care about' that we seek to 'lead our lives in light of', which 'set boundaries or perimeters to what we genuinely prefer' (p. 152). In other words, individuals' preferences reflect their subjective interests when they are consistent with the values they hold. The notion of 'perimeters of values' highlights

that, in any given situation, there may not be a unique preference over options that is consistent with an agent's values but a set of multiple preferences.

This does not constitute a commitment to *value fulfilment theory* as the correct substantive theory of well-being (on this see, e.g. Raibley, 2010; Tiberius, 2018), but rather an assumption that individuals form preferences in light of values that they hold and that the realisation of these values capture, but do not necessarily constitute, individuals' own good. This view of subjective interests is compatible with the critics' modelling of individuals' decision-making discussed in the previous section and in line with Thoma's (2021) position that an 'anti-paternalist' behavioural welfare economics ought to take 'underlying, more fundamental attitudes' to constitute subjective interest (p. 355), which is a view that (already) 'remains more or less implicit in the behavioural welfare economics literature' (p. 361).

The second criterion for permissible preference purification is that the correction of any given context-dependence must lead to greater realisation of individuals' values, as balanced by the evaluations they identify with at the moment of choice. This precludes analysts from disregarding an evaluation of options' attributes that an agent has settled on, within the perimeters of her values, in favour of another evaluation that is *also* within the perimeters of her values. In other words, for an agent's choice to constitute a 'mistake', the choice must fail to efficiently promote her subjective interests.

Drawing on these considerations, the following section discusses the permissibility of purifying individuals' preferences on the basis of different sources of context-dependence. Before proceeding, it is worth emphasising the scope of this argument. When arguing that context-dependent choice in a given decision setting can permissibly be purified, the claim that is made is that behavioural welfare economists are justified in holding a presumption in favour of one context-dependent choice over another being more welfare-relevant in light of an agent's subjective interests. This provides a partial justification for intervening in her decision to steer her in the direction of the option that would make her better off in light of her subjective interests. However, it is important to note that the claim 'an agent would be better off in light of her subjective interests were she to choose a particular option' – which is the type of claim this article is concerned with – does not imply the claim 'an agent would be better off in light of her subjective interests were she to choose a particular option *as a result of an intervention*'. Individuals do not only value the goods promoted by obtaining an option, but many times value being the authors of their own lives in the sense of making their own choices without undue external influence. A complete analysis of the welfare effects of particular paternalistic policies, which is beyond the scope of this article, should therefore 'consider the support for an intervention's goal along with the opposition to its method' (Arad & Rubinstein, 2018, p. 331).[3]

4. Three sources of context-dependence

This section proposes a classification of context-dependent choice based on the impact of contextual cues on different aspects of individuals' decision-making. The classification is based on a broadly folk-psychological representation of individuals' decision-making, implied in most behavioural welfare economic approaches and in the analyses of the programme's critics, as one where individuals: (i) make (or hold) evaluations of options' characteristics and consequences that constitute their evaluative perspective; (ii) hold beliefs about available options' characteristics and consequences; and (iii) proceed to act on the basis of these evaluations and beliefs when making a choice.

In light of this representation, I delineate three broad sources of context-dependence (without claiming to exhaust all possible sources): variation in individuals' evaluative perspective, false beliefs, and self-control failure. I argue that false beliefs and self-control failure provide grounds for permissible preference purification in *principle*; the following section argues that behavioural welfare economists have access to empirical procedures that facilitate permissible preference purification in *practice*.

4.A. Variation in individuals' evaluative perspective as a source of context-dependence

A first source of context-dependence is variation in individuals' evaluative perspective. With evaluative perspective, I mean an agent's assignment of *relative weight* to options' attributes. For example – to continue with the example of the cafeteria – an agent may ground her preference over what food item to consume on the values 'being healthy', 'enjoying pleasurable experiences', and 'being vegetarian'. In a given situation, she may give lexicographical priority to 'being vegetarian', while assigning relatively more weight to 'being healthy' over 'enjoying pleasurable experiences'.

The challenge for making decisive welfare judgements is that individuals' evaluative perspective may vary with context. Some of the most important findings in behavioural economics concern psychological mechanisms that can be seen to impact individuals' evaluative perspective – for example, loss aversion (Kahneman & Tversky, 1984). The main insight from the literature is that people do not only care about absolute levels of different attributes of options, but changes to these levels in relation to some reference point – typically current endowment (Camerer & Loewenstein, 2004). People tend to be more sensitive to losses than gains, in relation to a given reference point, and are often risk-seeking over losses while being risk-averse over gains. Since most decisions involve trade-offs between different dimensions of options, how the options are framed in terms of losses and gains in the different dimensions will many times affect individuals' evaluative perspective: 'the weight assigned to the different dimensions depends systematically on whether the chooser sees them as a gain or a loss' (Grüne-Yanoff, 2016, p. 468), because, as Kahneman et al. (1991) report based on a lab experiment, 'subjects are more sensitive to the dimension in which they are losing relative to their reference point' (p. 201).

Is purifying individuals' context-dependent choices due to variation in their evaluative perspective permissible? Not as long as individuals' evaluations are within the perimeters of their values, as doing so would fail to respect targeted individuals' subjective interests. As discussed in Section 2, individuals' true preferences may be context-dependent in instances of multi-attribute decision problems when the relative weight individuals assign to their different values are partially determined by the context or when individuals are unable to trade off their different values. The broad reason behind this is that individuals' values are often vague, and with that somewhat permissive in their success criteria, which allows for some wiggle-room in individuals' evaluations, and thereby preferences, as they are constructed in different contexts. As put by Engelen and Nys (2020), individuals' values tend to be 'broader, more general' in nature (p. 152), which allows for 'some leeway' in how they are cashed out in individuals' preferences over options (p. 153).

This does not, however, mean that 'anything goes' when it comes to context-dependent evaluations – individuals' values do constrain their evaluations. On the one hand, individuals may act in a way that is patently outside the perimeters of their values. For example, while the cafeteria visitor making a decision on the basis of 'being healthy', 'enjoying pleasurable experiences', and 'being vegetarian' can sometimes choose cake and sometimes apples, choosing beef jerky would categorically conflict with her commitment to vegetarianism, which may be absolute. Indeed, 'some values constrain others insofar as the standards they impose on action make it impossible to pursue other values in certain ways' (Tiberius, 2018, p. 78).

On the other hand, individuals' values impose limits on the 'leeway' in their evaluations if we move from appraising individual choices to *sequences of choice*. The cafeteria visitor above may permissibly choose cake on individual occasions, but at some point her value 'being healthy' will be decisively undermined by choosing unhealthy options (cf. DesRoches, 2020).

4.B. False beliefs as a source of context-dependence

A second source of context-dependence is false beliefs about options' characteristics and consequences. That is, what the attributes of the options are and the impact of those attributes on what the agent cares about. This includes not only individuals' beliefs about the options'

consequences abstractly, but individuals' vivid appreciation of what it would mean *for them* to experience those consequences (cf. Brandt, 1979, p. 70; Griffin, 1986, pp. 314–315n18). An important aspect of this source of context-dependence is therefore individuals' understanding (or accurate forecasts) of their interests and behaviours over time.

A large number of biases and reasoning limitations discussed in the behavioural economics literature are psychological mechanisms that can be seen to impact individuals' beliefs about options and their consequences. This includes findings related to the fact that people are generally poor at making decisions under risk due to faulty probability judgements. This is an artefact of people relying on heuristics that, while appropriate and useful in some instances, will lead them astray in other contexts. This includes people's use of the 'representativeness heuristic', the 'availability heuristic', and 'anchoring' (Tversky & Kahneman, 1974). All of these heuristics affect people's ability to correctly recognise the risks involved in particular choices, which may lead to false beliefs about options. This effect may be more or less present in decision settings depending on how the contexts are constructed.

An example of this type of context-dependence is discussed by Köszegi and Rabin (2007, 2008). We may observe an agent, Tina, making a bet on which side a coin will land next, in a situation where the coin has landed on heads five times in a row. If Tina places a bet on tails, even though it has lower payoff than a bet on heads, the most plausible explanation – given a minimal assumption that Tina prefers more money to less – is that she is subject to the gambler's fallacy. That is, Tina believes that, since the coin has landed on heads five times in a row, tails is 'due'. Of course, this is not consistent with objective probabilities. This is an instance of context-dependent choice – Tina's preferences are dependent on the welfare-irrelevant contextual factor of 'outcomes of previous coin flips'. The mechanism underlying the context-dependence is the representativeness heuristic, which, in this case, leads to *false beliefs* about the options.

Beyond psychological mechanisms affecting individuals' risk judgements are findings that individuals systematically mispredict their future interests. As noted by Rabin (1998): 'Even when [people] correctly perceive the physical consequences of their decisions' – and the likelihood of these consequences materialising – 'people systematically misperceive the well-being they derive from such outcomes' (p. 33). This includes 'hedonic mispredictions'. People tend to remember extremes of pain and pleasure rather than the average; people tend to focus on the end of an episode in estimating their experience of the episode as a whole; people also tend to neglect the duration of an episode when retrospectively evaluating it (Kahneman, 1994). Since people often predict utility of future experiences by recollecting utility from comparable past experiences, this type of biased estimates leads to forecasting errors.

This type of misprediction of utility also includes 'projection bias', where individuals project their current preferences into future periods, which may correspond poorly with their actual future utility (Loewenstein et al., 2003). It further includes broader failures to appreciate adaption to new circumstances. This is of special interest in the health domain as individuals' failure to consider adaptation lead them to mispredict the impact of disabilities on their quality of life (Ubel et al., 2005).

A possible example of this type of context-dependence is Read and van Leeuwen's (1998) field experiment. In the experiment, office workers were asked to make an 'advance choice' of a snack (which could be something healthy, like an apple, or something unhealthy, like a Mars bar) that they would receive at a designated time one week later; they were then asked to make an 'immediate choice' over the same snacks at the designated time in the future. The participants were assumed to be either 'satiated' (making their choice directly after lunch) or 'hungry' (making their choice in the late afternoon) when making both their advance choice and their immediate choice. Read and van Leeuwen found that the participants' advance choices – for a given time of receipt of the item the following week – were affected by the time of day at which the choice was made: the group that was hungry when making the advance choice chose unhealthy snacks to a significantly higher degree than the group that was satiated.

Read and van Leeuwen's (1998) favoured explanation of the observed context-dependence is based on projection bias. Participants in the experiment are making choices for their future selves. They can then be seen to be making predictions about their future preferences and making choices that they believe best satisfy those future preferences. What causes the observed context-dependence is that people are subject to an 'intrapersonal empathy gap', 'caused by the imperfect ability of decision-makers to project themselves into circumstances different than those in which they find themselves' (p. 202). To predict their future preferences, people tend to simply project their current preferences into future periods. This becomes problematic when their preferences differ significantly over contexts due to environmental factors, such as experienced hunger, as it leads to mispredictions – in other words, choices based on *false beliefs* about their future selves.

Contrary to context-dependence due to variation in individuals' evaluative perspective within the perimeters of their values, context-dependence due to false beliefs can permissibly be purified. On the one hand, this is the case since it does not threaten the subjectivity of individuals' true preferences: belief-based preference purification targets non-evaluative facts about options' attributes grounding individuals' preferences, not targeted individuals' evaluations of those attributes. On the other hand, this is the case since a preference based on false beliefs is 'not actually for the option as it is but rather for the option as it is falsely imagined to be' (Sobel, 2009, p. 345) and will therefore not (efficiently) advance the agent's subjective interests (barring sheer luck). To take an obvious example:

> Suppose somebody chooses a glass of juice over a glass of water without knowing that the former contains deadly poison. From this fact we obviously cannot infer that he really prefers to drink the poison, or that drinking the poison is in his real interest. (Harsanyi, 1992, p. 702)

This point is worth emphasising in relation to a case that Infante et al. (2016b) take to be an example of context-dependent choice, but where 'the definition of a person's true preferences or best interests is fairly uncontroversial' (p. 4). The case in question concerns consumers making a choice among competing suppliers in retail energy markets offering exactly the same product but priced according to different tariffs. Due to the obscure ways tariffs are described, consumers may end up not choosing the supplier with the lowest final price. Infante and co-authors suggest that 'representing such choices as mistakes, defined relative to "true" preferences for low prices, may be a reasonable modelling strategy' (p. 4). They quickly add, however, that 'the assumption that is taken to be uncontroversial in defining mistakes equates the consumer's *subjective* ranking of options (alternative tariffs) with an *objective* ranking (in inverse order of their prices) that is independent of the consumer's perception or judgements' (p. 4).

This is not necessarily correct. As Infante et al. (2016b) note, that 'consumers have an underlying preference for paying less rather than more' (p. 4) is an *assumption*. It is an assumption about individuals' subjective interests that targeted individuals can accept or reject. Importantly, it is an assumption that can be tested by, for example, asking targeted individuals what goal(s) they aim to achieve with the particular interaction. There is therefore nothing 'objective' about this ranking, beyond the trivial sense that it is a goal that presumably many (if not all) individuals share in the domain of energy supply. Again, belief-based preference purification preserves the subjectivity of true preferences.

4.C. Self-control failure as a source of context-dependence

A third source of context-dependence is variability in individuals' ability to act in accordance with their pre-settled and continuously endorsed preferences; more specifically, failed attempts by individuals to resist present temptation and implement preferences they have already formed and that they identify with.

Psychological mechanisms related to this class of 'self-control failures' tend to revolve around discussions of time-inconsistent preferences. Individuals are taken to be 'present biased' in that they

discount 'near-term incremental delays in well-being more severely than […] distant future incremental delays' (Rabin, 2002, p. 668). As put by DellaVigna (2009), 'when evaluating outcomes in the distant future, individuals are patient and make plans to exercise, stop smoking, and look for a better job'; however, 'as the future gets near, the discounting gets steep, and the individuals engage in binge eating, light another (last) cigarette, and stay put on their job' (p. 318). In other words, there is an inconsistency between individuals' far-sighted plans of action, or 'resolutions', and their revealed preferences at the moment of choice.

Behavioural welfare economists tend to assign normative authority to individuals' resolutions when there is an observed conflict with their revealed preferences at the moment of choice (Whitman & Rizzo, 2015). However, individuals' failure to act on their resolutions need not be a mistake. Take the case of Tom, as discussed by Thaler and Sunstein (2008):

> Tom is on a diet and agrees to go out on a business dinner, thinking that he will be able to limit himself to one glass of wine and no dessert. But the host orders a second bottle of wine and the waiter brings by the dessert cart, and all bets are off. (p. 42)

The inconsistency between Tom's resolution to limit his consumption of alcohol and dessert, and his revealed preference at the moment of choice to indulge his desires, could be due to reasonable context-dependence in his evaluative perspective, as discussed in Section 4.A: Tom could judge that the current instance of choice is a moment for spontaneity, not resoluteness (cf. Sugden, 2018b), while staying committed to fulfilling his goal of losing weight. Alternatively, the inconsistency could reflect a change of mind about the value of the resolution – Tom may no longer recognise dieting as genuinely reflecting his better judgement. More generally, recent empirical research highlights significant heterogeneity in individuals' attitudes toward acting impulsively (Ghoniem & Hofmann, 2021; Grubiak et al., 2022) – individuals do not always favour promoting their long-term goals over acting on present temptations.

In order for self-control failure to be permissibly corrected something more is then needed: an acknowledgement by the agent herself, at the moment of choice, that she is acting against her better judgement. That is, that Tom feels the urge, based on contextual cues, to act on immediate gratification; attempts to resist the temptation in order to act on pre-settled preferences he identifies with; but ultimately fails. This point is recognised by Sugden in a number of commentaries related to the critique of the inner rational agent (Sugden, 2017, 2018a, 2018b), who grants that this can, in some instances, substantiate the libertarian paternalist subjectivist mantra of making 'choosers better off, as judged by themselves'.

Context-dependence due to *self-acknowledged* self-control failure can therefore permissibly be purified. Individuals, in these situations, fail to act on evaluations of options' attributes, which they identify with at the moment of choice, that are within the perimeters of their values. Acting on temptations, when in self-acknowledged conflict with endorsed evaluations, is clearly alienating for individuals as it is in opposition to their valuational systems and can therefore not be in their subjective interest. Of course, this is not to say that individuals' resolutions are beyond criticism: they can still be based on false beliefs and thereby fail to promote their subjective interests, as discussed in Section 4.B.

5. Establishing the source of context-dependence

The previous section provided a broad classification of context-dependent choice based on the impact of contextual cues, mediated by different psychological mechanisms, on different aspects of individuals' decision-making. Locating the effects of cognitive biases and reasoning impairments in individuals' decision-making is important, as it allows for a more systematic discussion of the permissibility of purifying preferences based on given psychological mechanisms.

The permissibility of correcting context-dependence depends on its source in individuals' decision-making: correcting context-dependent choices due to variation in individuals' evaluative

perspective is impermissible as long as the evaluations are within the perimeters of targeted individuals' values; however, in instances where the source of the context-dependence is either false beliefs or self-control failure, context-dependent choices can permissibly be purified with regard to individuals' subjective interests.

This is all to say that there are conceptual resources available to behavioural welfare economists to permissibly purify individuals' context-dependent choices in some instances: there is nothing methodologically or ontologically suspicious in correcting context-dependence by appealing to true preferences in *principle*. However, critics of behavioural welfare economics may still emphasise epistemological concerns associated with purifying individuals' context-dependent choices in *practice*.

A salient challenge for permissibly correcting context-dependent choice is one of 'underdetermination' – choice patterns will often be compatible with many different models, and interpretations of these models, of individuals' reasoning behind their choices (Manzini & Mariotti, 2014). The previously discussed example of Read and van Leeuwen's (1998) field experiment helps illustrate this concern.

The explanation proposed by Read and van Leeuwen (1998) of why the two groups (differing in their experienced hunger when making the choice) make different choices of what snack to consume in the future is that they are to varying degrees subject to an 'intrapersonal empathy gap', caused by the fact that 'it is inordinately difficult for us to imagine what it is like to be in a different visceral state than the one we are currently in' (p. 190). The underlying assumption for Read and van Leeuwen is that individuals in this situation are trying to predict, and make choices that best satisfy, their future preference. On this account of individuals' decision-making, the context-dependence can be attributed to the source of *false beliefs*: in the satiated context, the agent acts on false beliefs about the preferences of her future self; in the hungry context, the agent acts on beliefs about her future preferences that to a greater degree correspond with what will be the case.

Infante et al.'s (2016b) proposed explanation of the observed context-dependence is instead evaluation-based: 'the hungrier you feel, the more attention you give to cues that are directed towards the satisfaction of hunger'; therefore, 'the hunger-satisfying properties of the Mars bars are perceived more vividly in the late afternoon, irrespective of when it will actually be eaten' (p. 3), which leads to variation in *evaluations* of the options between the two contexts.

The context-dependence can also be construed as being due to variability in self-control. It may be that the satiated context captures the agent's endorsed preferences, but that the agent falls for the temptation of the tasty Mars bar when hungry – even if the decision concerns future consumption. This account can further be filled in to cover that the agent acknowledges acting against her better judgement in the moment of choice and tries, but ultimately fails, to resist the temptation. On this account of how the agent relates to the decision, the context-dependence can be attributed to the source of *self-control failure*.

This is then an instance of context-dependent choice which is consistent with all three sources explaining the context-dependence. It is '*as if* individuals seek to predict and satisfy their future preferences, but sometimes fail due to an inability to imagine their future preferences'; it is also '*as if* individuals make all-things-considered judgements about the value of the options, but that the relative importance of options' attributes are partially determined by contextual salience'; and, lastly, it is '*as if* individuals hold endorsed preferences over the options, but that they sometimes fail (not for want of trying) to act on these preferences due to contextual temptations'.

Only if individuals relate to the decision in the first or third sense is preference purification permissible. Moreover, the different accounts would prescribe correcting individuals' choices in different directions: if the source is false beliefs, the choice made when satiated should be purified; if the source is self-control failure, the choice made when hungry should be purified. It is therefore essential for the behavioural welfare economic project that credible evidence about how individuals *actually* interact with particular decisions is collected in order to discriminate between the different sources.

Fortunately, empirical procedures are available to behavioural welfare economists that can support credible attributions to sources in practice. These procedures will be familiar to welfare economists, but have not generally been used to discern sources of context-dependence in given decision settings – in particular, whether the observed context-dependence is due to false beliefs or self-control failure.

One central question to ask when establishing the source of a given context-dependent choice is whether the choice pattern is caused by false beliefs. An important consideration in this regard is whether any belief identified as faulty actually plays a causal role in individuals' decisions. For example, it is not enough to show that individuals hold a false belief about their future preference over snacks in Read and van Leeuwen's (1998) field experiment in order to permissibly correct that choice. If individuals do not base their choice of snack to consume in the future on beliefs about future preferences, the false belief cannot generate the observed context-dependence and thereby justify purification of the choice.

'Think aloud' techniques provide an informative method for inquiring into individuals' reasoning process, their concerns, and their beliefs about options (Ericsson & Simon, 1993). The core idea behind the method is simply to ask respondents to verbalise their thought process as they are completing choice tasks. The method has primarily been used in cognitive psychology and educational science, but has become increasingly popular in health economics. Ryan et al. (2009) use think aloud to get a better picture of the underlying reasons for respondents' ill-formed preferences in a health state valuation study. They find, as is common in these studies, that respondents infer additional information that is inconsistent with the actual outcomes of options and misunderstand attributes of options. That is, they find evidence of false beliefs driving individuals' decisions. Such information can be used to exclude certain choice contexts from the domain of welfare-relevant choices and construct, or privilege, choice contexts that do not induce the formation of preferences based on false beliefs.

Comprehension tests of verifiable facts about options' outcomes or understanding of concepts central for making informed decisions constitute another method of discerning the source of context-dependent choice. Survey-based comprehension tests have been used to highlight consumers' poor understanding of health insurance coverage (Loewenstein et al., 2013) and low level of financial literacy (Lusardi & Mitchell, 2014). Applied to specific decision settings, such tests can support the exclusion of certain choice contexts from the welfare-relevant domain in favour of simplified choice contexts or contexts with relevant educational interventions.

Another question to ask when establishing the source of a given context-dependent choice is whether the choice pattern is caused by self-control failure. Experience sampling, which overlaps significantly with ecological momentary assessment (Rabasco & Adnover, 2022), can be used to support attributions in this regard. The basic idea behind experience sampling is to 'ping' individuals to complete concise self-reports (e.g. on a smartphone) about what they are doing, thinking, and feeling at various occasions in their everyday lives (Larson & Csikszentmihalyi, 1983/2014; Pejovic et al., 2016). These features – namely, at-the-moment and real-life sampling – makes the method well-suited for studies of self-control failure, since the relevant benchmark is self-acknowledged failures to resist temptations, recognised at the moment of choice (as discussed in Section 4.C).

Hofmann et al. (2012) use experience sampling to study the prevalence of self-control failure. At random occasions throughout a week, Hofmann and co-authors pinged respondents to report on (i) whether they were currently experiencing, or had just experienced, a desire; (ii) the content and strength of this desire; (iii) whether they had attempted to resist the desire; (iv) whether they had enacted the desire; (v) whether the desire conflicted with a personal goal; and (vi) the content and strength of this goal. They found that conflicts between individuals' goals and desires were common and that individuals enacted desires even though they had attempted to resist them in a significant number of cases. Future research may use event-contingent sampling (when specific events occur), rather than the signal-contingent sampling (at random moments) of Hofmann and

co-authors, to study self-control failure in relation to different choice contexts in given decision settings, in order to exclude, from the welfare-relevant domain, contexts where self-control failure is prevalent in favour of contexts where this is not the case.

Experience sampling can additionally be used to estimate self-control failure by providing data on 'spoiled pleasure'. Hofmann et al. (2013) find evidence of respondents in an experience sampling study receiving considerably smaller gains in momentary happiness from enacting desires conflicting with long-term goals compared to non-conflicting desires, mediated by emotions of guilt and reduced pride when failing to resist goal-conflicting desires. If further corroborated, such data on spoiled pleasure can be used to support arbitration between different choice contexts on grounds of self-control failure.

In sum, behavioural welfare economists have access to empirical procedures that can be used to make credible attributions of context-dependent choice to particular sources in individuals' decision-making. It is therefore not correct to say, as is sometimes suggested by critics of behavioural welfare economics, that 'we are given no independent criterion by which errors can be identified' in behavioural welfare economics (Sugden, 2017, p. 117). As discussed in the previous section, there are viable independent standards of correctness that can be invoked to permissibly purify individuals' preferences (false beliefs, self-control failure); and, as discussed in this section, these standards can be operationalised to assess the welfare-relevance of individuals' preferences in different contexts.

This is not to deny that there are challenges associated with the programme proposed in this article. For one, convincingly establishing the source of context-dependent choices will plausibly require significantly more use of non-choice data (e.g. self-reports on cognition or affect, comprehension tests) than what is commonly used in welfare economics. Economists have a historical aversion to using non-choice data since 'talk is cheap', and one might worry about reactivity on the part of participants due to social desirability bias or altered reasoning processes as a result of the empirical procedures. Proponents of these procedures are, of course, aware of these challenges and there are methods to allay some of these concerns (see, e.g. discussion in Ericsson & Simon, 1993; Scollon et al., 2003). For example, comprehension tests of verifiable facts and concepts can be incentivised (as in Ambuehl et al., 2022), and triangulation (e.g. using both self-reports on mental processes and momentary happiness data, in the case of self-control failure) can be used to strengthen the credibility of the evidence.

Related to this is a point about costliness. If required to carry out additional empirical procedures (e.g. the ones discussed above) before purifying individuals' preferences, behavioural welfare economic analysis and related policymaking will become more expensive and time-consuming. This, however, seems inevitable if taking serious true preferences as a *subjective* standard of welfare. Since context-dependent choice is in itself not a mistake, evidence of false beliefs or self-control failure generating the context-dependence is needed to permissibly purify individuals' preferences. *How much* evidence is needed for analysts or policymakers to justifiably correct individuals' choices I leave unanswered. Opinions differ in the literature on the burden of proof when intervening in individuals' decisions (cf. Engelen & Nys, 2020, p. 154; Thoma, 2021, p. 360), which may also depend on the degree of coerciveness of given policy interventions (Hausman, 2022a; Rizzo & Whitman, 2021; Sunstein & Thaler, 2003).

6. Conclusion

What are the implications of the critique of the inner rational agent for behavioural welfare economics? In this article, I have argued that the critique does not support a wholesale shift away from the use of true preferences as evaluative standard in normative economics. Rather, the critique implies that behavioural welfare economists need to inquire deeper into, and distinguish between, different sources of observed context-dependent choice in individuals' decision-making. I have argued that context-dependent choice due to false beliefs or self-control failure can permissibly be purified on viable subjectivist grounds and, further, that empirical procedures are available to

behavioural welfare economists to substantiate such attributions. This will plausibly require behavioural welfare economists to make more use of non-choice data than what has traditionally been the case, but this seems inevitable given a genuine commitment to true preferences as a subjective concept and the inherent limitations of neoclassical consistency of preferences as a measure of decision quality. The upshot is that behavioural welfare economics has access to conceptual and methodological resources to facilitate permissible purification of individuals' preferences, and with that decisive welfare judgements, in a number of situations.

This article has followed the norm in the behavioural welfare economics literature of taking a synchronic perspective on individuals' choices: analysing single instances of choice and taking individuals' values as settled. To a first approximation, this perspective is plausible: we can expect welfare assessments of individuals' sequences of choice to supervene on single choices they make and we can expect individuals' values to be at least 'semi-settled' (Hausman, 2022b, p. 346). However, interesting additional considerations and avenues for future research open when these assumptions are questioned and when the perspective is extended to analyse individuals' choices diachronically.

On the one hand, that individuals' values are sometimes (if not in general) 'vague', in the sense that the values have vague success criteria, opens for dynamic choice problems as discussed in a philosophical literature centred around Quinn's (1990) 'puzzle of the self-torturer' (see, e.g. Tenenbaum, 2020; Tenenbaum & Raffman, 2012). An agent may be in the (unfortunate) situation that even though she chooses within the perimeters of her values on any given occasion, she may yet end up frustrating her values over a sequence of choice. Future work may explore the permissibility of purifying, and if so how to permissibly purify, such a sequence of choice.

On the other hand, individuals' values may change over time and as a consequence of making certain decisions. So-called transformative experiences is an especially interesting type of such value change (Paul, 2014). Transformative experiences are such that, in making a particular choice, an agent's epistemic perspective is transformed in a way that her 'core personal preferences are significantly changed, leading to a significant change in how the post-change [self] would evaluate the act' (p. 51). Standard examples concern 'major life events' such as choosing a career, undergoing a major surgery, or becoming a parent (p. 42). The possibility of value change, such as transformative experiences, complicates things for analysts seeking to make welfare judgements based on individuals' subjective interests. In this article, the assumption has been that individuals' values remain stable over the course of making a choice, which means that there is a clear in sense in which correcting for false beliefs and self-control failure will leave individuals better off as judged by their subjective interests both before and after the decision. But if individuals' values change significantly over the course of making a choice, we can imagine special cases where an agent acts against her subjective interests at the moment of making a decision (possibly due to false beliefs) and yet end up with a choice that is in line with her subjective interests *after* the choice. The possibility of such cases does not challenge the general presumption, expressed in this article, that choices due to false beliefs and self-control failure should be excluded from the domain of welfare-relevant choices; however, it does call for additional considerations in the particular decision settings where value change can be expected. If individuals' subjective interests change as a consequence of making a choice, should the welfare judgement about the choice be based on targeted individuals' subjective interests before or after the choice? Whether there are more or less 'correct', or reasonable, ways for individuals to approach these particular situations and trade off the interests of their changing selves (see, e.g. Bykvist, 2006, 2022; Pettigrew, 2019), that allow for paternalistic policy-guidance (Paul & Sunstein, 2019; Pettigrew, 2023), is a topic that warrants further attention.

Notes

1. Since context-dependence is understood as being due to *irrelevant* features of decision settings, it does not concern choices affected by features of decision settings that constitute *good reasons* for individuals altering

their choices. For example, redescriptions of options, or the inclusion/exclusion of some options, providing *more* information about available options, or where alternatives-not-chosen are relevant to choosers (see, e.g. Sen, 1993 on these points).

2. Rizzo and Whitman (2020) provide two additional justifications for why context-dependent choice may be reasonable: (i) individuals are in the process of 'discovering' their underlying preferences and (ii) individuals are 'economising' on the cognitive efforts that would be needed to form consistent preferences. I do not discuss these justifications since they do not transfer to 'ideal reasoning circumstances' where individuals have 'unlimited cognitive abilities' and 'complete information'. Infante et al. (2016b) provide an additional justification for why individuals' true preferences may be context-dependent: individuals may individuate the options differently than the analyst. I do not discuss this justification since it does not constitute a principled justification for context-dependent true preferences, but highlights a practical problem for analysts in correctly identifying the scope of individuals' (reasonable) concerns.
3. Much more can be, and has been, said about the implications for paternalistic policymaking, such as nudging, of individuals valuing making their own choices without undue external influence. Even though many individuals value making their own choices (including mistaken ones, in light of their subjective interests), it is important to recognise that individuals have limited (mental) 'bandwidth' (Mullainathan & Shafir, 2013); appropriately exercising their agency will therefore plausibly entail delegating some choices, or aspects of choosing, to others – individuals may sometimes 'choose not to choose' (Sunstein, 2014). Whether a particular intervention is permissible in light of individuals' values on balance (concerning both outcomes and the choice process itself) will plausibly depend on considerations such as the centrality of the choice for an agent's identity, the complexity of the choice, the already existing (private) external influences prior to the intervention, and the channel by which the intervention influences an agent's decision-making. Evidence relating to these, and similar, considerations should plausibly figure as part of a complete welfare analysis of *particular* paternalistic policies in given decision settings.

Acknowledgements

Earlier versions of the paper have been presented at ECAP10 (2020), Utrecht; OZSW Annual Conference (2020), Tilburg; EIPE PhD seminar (2021), Rotterdam; Conference on Preferences, Commitments, and Choice (2021), Zürich; and INEM 15th Biennial Conference (2021), Arizona. I thank audience members for many useful comments. I wish to give special thanks to Lukas Beck, Constanze Binder, Bart Engelen, James Grayot, Ivan Moscati, Johanna Thoma, and Jack Vromen for helpful feedback and discussion of earlier versions of the paper. I am grateful for comments by two anonymous referees for this journal that helped improve the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project has received funding from the Erasmus Initiative 'Smarter Choices for Better Health '.

Notes on contributor

Måns Abrahamson is a PhD candidate at the Erasmus School of Philosophy, Erasmus University Rotterdam. His fields of interest include philosophy of economics, value theory, and public policy. He is particularly interested in the conceptualisation and use of 'laundered preferences' in economic evaluations.

References

Ambuehl, S., Bernheim, B. D., & Lusardi, A. (2022). Evaluating deliberative competence: A simple method with an application to financial choice. *American Economic Review*, *112*(11), 3584–3626. https://doi.org/10.1257/aer.20210290

Arad, A., & Rubinstein, A. (2018). The people's perspective on libertarian-paternalistic policies. *The Journal of Law and Economics*, *61*(2), 311–333. https://doi.org/10.1086/698608

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. https://doi.org/10.1017/bca.2016.5

Bernheim, B. D. (2021). In defense of behavioral welfare economics. *Journal of Economic Methodology*, *28*(4), 385–400. https://doi.org/10.1080/1350178X.2021.1988133

Bernheim, B. D., & Rangel, A. (2007). Toward choice-theoretic foundations for behavioural welfare economics. *American Economic Review*, *97*(2), 464–470. https://doi.org/10.1257/aer.97.2.464

Bernheim, B. D., & Rangel, A. (2009). Beyond revealed preferences: Choice-theoretic foundations for behavioural welfare economics. *Quarterly Journal of Economics*, *124*(1), 51–104. https://doi.org/10.1162/qjec.2009.124.1.51

Brandt, R. B. (1979). *A theory of the good and the right*. Clarendon Press.

Buchanan, J. M. (1999). Natural and artifactual man. In J. M. Buchanan (Ed.), *The collected works of James M. Buchanan: Vol 1. The logical foundations of constitutional liberty* (pp. 246–259). Liberty Fund (Original worked published 1979)

Bykvist, K. (2006). Prudence for changing selves. *Utilitas*, *18*(3), 264–283. https://doi.org/10.1017/S0953820806002032

Bykvist, K. (2022). Well-being and changing attitudes across time. *Ethical Theory and Moral Practice*, 1–15. Advance online publication.

Camerer, C. F., & Loewenstein, G. (2004). Behavioral economics: Past, present, future. In C. F. Camerer, G. Loewenstein, & M. Rabin (Eds.), *Advances in behavioral economics* (pp. 3–51). Princeton University Press.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, *47*(2), 315–372. https://doi.org/10.1257/jel.47.2.315

DesRoches, T. (2020). Value commitment, resolute choice, and the normative foundations of behavioural welfare economics. *Journal of Applied Philosophy*, *37*(4), 562–577. https://doi.org/10.1111/japp.12418

Dold, M. F. (2018). Back to Buchanan? Explorations of welfare and subjectivism in behavioural economics. *Journal of Economic Methodology*, *25*(2), 160–178. https://doi.org/10.1080/1350178X.2017.1421770

Engelen, B., & Nys, T. (2020). Nudging and autonomy: Analyzing and alleviating the worries. *Review of Philosophy and Psychology*, *11*(1), 137–156. https://doi.org/10.1007/s13164-019-00450-z

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). The MIT Press.

Ghoniem, A., & Hofmann, W. (2021). When impulsive behaviours do not equal self-control failures: The (added) value of temptation enactments. *European Journal of Personality*, *35*(2), 267–288. https://doi.org/10.1002/per.2280

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

Grubiak, K. P., Isoni, A., Sugden, R., Wang, M., & Zheng, J. (2022). Taking the new year's resolution test seriously: Eliciting individuals' judgements about self-control and spontaneity. *Behavioural Public Policy*, 1–23. Advance online publication.

Grüne-Yanoff, T. (2016). Why behavioural policy needs mechanistic evidence. *Economics and Philosophy*, *32*(3), 463–483. https://doi.org/10.1017/S0266267115000425

Harsanyi, J. C. (1992). Game and decision theoretic models in ethics. In R. J. Aumann & S. Hart (Eds.), *Handbook of game theory: Volume 1* (pp. 669–707). Elsevier.

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

Hausman, D. M. (2022a). Banishing the inner econ and justifying paternalistic nudges. *Behavioural Public Policy*, 1–12. Advance online publication.

Hausman, D. M. (2022b). Enhancing welfare without a theory of welfare. *Behavioural Public Policy*, *6*(3), 342–357. https://doi.org/10.1017/bpp.2019.34

Hédoin, C. (2017). Normative economics and paternalism: The problem with the preference-satisfaction account of welfare. *Constitutional Political Economy*, *28*(3), 286–310. https://doi.org/10.1007/s10602-016-9227-5

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. https://doi.org/10.1177/1745691617702496

Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012). Everyday temptations: An experience sampling study on desire, conflict, and self-control. *Journal of Personality and Social Psychology*, *102*(6), 1318–1335. https://doi.org/10.1037/a0026545

Hofmann, W., Kotabe, H., & Luchmann, M. (2013). The spoiled pleasure of giving in to temptation. *Motivation and Emotion*, *37*(4), 733–742. https://doi.org/10.1007/s11031-013-9355-4

Infante, G., Lecouteux, G., & Sugden, R. (2016a). "On the econ within": A reply to Daniel Hausman. *Journal of Economic Methodology*, *23*(1), 33–37. https://doi.org/10.1080/1350178X.2015.1070526

Infante, G., Lecouteux, G., & Sugden, R. (2016b). 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. https://doi.org/10.1080/1350178X.2015.1070527

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional Theoretical Economics*, *150*(1), 18–36.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, *5*(1), 193–206. https://doi.org/10.1257/jep.5.1.193

Kahneman, D., & Tversky, A. (1984). Choice, values, and frames. *American Psychologist*, *39*(4), 341–350. https://doi.org/10.1037/0003-066X.39.4.341

Köszegi, B., & Rabin, M. (2007). Mistakes in choice-based welfare analysis. *American Economic Review*, *97*(2), 477–481. https://doi.org/10.1257/aer.97.2.477

Köszegi, B., & Rabin, M. (2008). Choices, situations, and happiness. *Journal of Public Economics*, *92*(8–9), 1821–1832. https://doi.org/10.1016/j.jpubeco.2008.03.010

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In M. Csikszentmihalyi (Ed.), *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi* (pp. 21–34). Springer. (Original work published 1983)

Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference*. Cambridge University Press.
Lipman, S. A., Brouwer, W. B. F., & Attema, A. E. (2019). The corrective approach: Policy implications of recent developments in QALY measurement based on prospect theory. *Value in Health*, *22*(7), 816–821. https://doi.org/10.1016/j.jval.2019.01.013
Loewenstein, G., Friedman, J. Y., McGill, B., Ahmad, S., Linck, S., Sinkula, S., Beshears, J., Choice, J. J., Kolstad, J., Laibson, D., Madrian, B. C., List, J. A., & Volpp, K. G. (2013). Consumers' misunderstanding of health insurance. *Journal of Health Economics*, *32*(5), 850–862. https://doi.org/10.1016/j.jhealeco.2013.04.004
Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Projection bias in predicting future utility. *The Quarterly Journal of Economics*, *118*(4), 1209–1248. https://doi.org/10.1162/003355303322552784
Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, *52*(1), 5–44. https://doi.org/10.1257/jel.52.1.5
Manzini, P., & Mariotti, M. (2014). Welfare economics and bounded rationality: The case for model-based approaches. *Journal of Economic Methodology*, *21*(4), 343–360. https://doi.org/10.1080/1350178X.2014.965909
Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much*. Times Books.
North, A. C., Hargreaves, D. J., & McKendrick, J. (1997). In-store music affects product choice. *Nature*, *390*(6656), 132. https://doi.org/10.1038/36484
Paul, L. A. (2014). *Transformative experiences*. Oxford University Press.
Paul, L. A., & Sunstein, C. R. (2019). "*As judged by themselves": Transformative experiences and endogenous preferences* [Unpublished manuscript]. Department of Philosophy, Yale University.
Pejovic, V., Lathia, N., Mascolo, C., & Musolesi, M. (2016). Mobile-based experience sampling for behaviour research. In M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, & A. Košir (Eds.), *Emotions and personality in personalized services* (pp. 141–161). Springer.
Pettigrew, R. (2019). *Choosing for changing selves*. Oxford University Press.
Pettigrew, R. (2023). Nudging for changing selves. *Synthese*, 1–21. Advance online publication.
Pinto-Prades, J.-L., & Abellan-Perpiñan, J.-M. (2012). When normative and descriptive diverge: How to bridge the difference. *Social Choice & Welfare*, *38*(4), 569–584. https://doi.org/10.1007/s00355-012-0655-5
Quinn, W. S. (1990). The puzzle of the self-torturer. *Philosophical Studies*, *59*(1), 79–90. https://doi.org/10.1007/BF00368392
Rabasco, A., & Adnover, M. (2022). Ecological momentary assessment. In D. McKay (Ed.), *Comprehensive clinical psychology: Vol 3. Research and methods* (2nd ed., pp. 83–90). Elsevier.
Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature*, *31*(1), 11–46.
Rabin, M. (2002). A perspective on psychology and economics. *European Economic Review*, *46*(4–5), 657–685. https://doi.org/10.1016/S0014-2921(01)00207-0
Raibley, J. (2010). Well-being and the priority of values. *Social Theory and Practice*, *36*(4), 593–620. https://doi.org/10.5840/soctheorpract201036432
Read, D., & van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes*, *76*(2), 189–205. https://doi.org/10.1006/obhd.1998.2803
Rizzo, M. J., & Whitman, D. G. (2020). *Escaping paternalism: Rationality, behavioral economics, and public policy*. Cambridge University Press.
Rizzo, M. J., & Whitman, D. G. (2021). The unsolved Hayekian knowledge problem in behavioral economics. *Behavioural Public Policy*, 1–13. Advance online publication.
Rubinstein, A., & Salant, Y. (2012). Eliciting welfare preferences from behavioural data sets. *Review of Economic Studies*, *79*(1), 375–387. https://doi.org/10.1093/restud/rdr024
Ryan, M., Watson, V., & Entwistle, V. (2009). Rationalising the 'irrational': A think aloud study of discrete experiment responses. *Health Economics*, *18*(3), 321–336. https://doi.org/10.1002/hec.1369
Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging: An overview. *Philosophy Compass*, *15*(4), e12658. https://doi.org/10.1111/phc3.12658
Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, *4*(1), 5–34. https://doi.org/10.1023/A:1023605205115
Sen, A. (1993). Internal consistency of choice. *Econometrica*, *61*(3), 495–521. https://doi.org/10.2307/2951715
Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, *29*(3), 281–295. https://doi.org/10.1177/002224379202900301
Sobel, D. (2009). Subjectivism and idealization. *Ethics*, *119*(2), 336–352. https://doi.org/10.1086/596459
Sugden, R. (2004). The opportunity criterion: Consumer sovereignty without the assumption of coherent preferences. *American Economic Review*, *94*(4), 1014–1033. https://doi.org/10.1257/0002828042002714
Sugden, R. (2015). Looking for a psychology for the inner rational agent. *Social Theory and Practice*, *41*(4), 579–598. https://doi.org/10.5840/soctheorpract201541432
Sugden, R. (2017). Do people really want to be nudged towards healthy lifestyles? *International Review of Economics*, *64*(2), 113–123. https://doi.org/10.1007/s12232-016-0264-1
Sugden, R. (2018a). 'Better off, as judged by themselves': A reply to Cass Sunstein. *International Review of Economics*, *65*(1), 9–13. https://doi.org/10.1007/s12232-017-0281-8

Sugden, R. (2018b). *The community of advantage: A behavioural economist's defence of the market*. Oxford University Press.
Sunstein, C. R. (2014). Choosing not to choose. *Duke Law Journal*, *64*(1), 1–52.
Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review*, *70*(4), 1159–1202. https://doi.org/10.2307/1600573
Tenenbaum, S. (2020). *Rational powers in action: Instrumental rationality and extended agency*. Oxford University Press.
Tenenbaum, S., & Raffman, D. (2012). Vague projects and the puzzle of the self-torturer. *Ethics*, *123*(1), 86–112. https://doi.org/10.1086/667836
Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
Thoma, J. (2021). On the possibility of an anti-paternalist behavioural welfare economics. *Journal of Economic Methodology*, *28*(4), 350–363. https://doi.org/10.1080/1350178X.2021.1972128
Tiberius, V. (2018). *Well-being as value fulfillment: How we can help each other to live well*. Oxford University Press.
Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics & biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683
Ubel, P. A., Loewenstein, G., & Jepson, C. (2005). Disability and sunshine: Can hedonic predictions be improved by drawing attention to focusing illusions or emotional adaption? *Journal of Experimental Psychology: Applied*, *11*(2), 111–123. https://doi.org/10.1037/1076-898X.11.2.111
Whitman, D. G., & Rizzo, M. J. (2015). The problematic welfare standards of behavioral paternalism. *Review of Philosophy and Psychology*, *6*(3), 409–425. https://doi.org/10.1007/s13164-015-0244-5