April 2023

iRisk WORKING PAPER SERIES

2023-iRisk-03

Inferring Welfare from Observed Choices: An Axiomatic Approach

Guilhem Lecouteux
Université Côte d'Azur, CNRS, GREDEG UMR 7321, 250 rue Albert Einstein, 06560 Valbonne, France

Ivan Mitrouchev
IESEG School of Management, Univ. Lille, CNRS, UMR 9221 - LEM - Lille Economie Management, F-59000 Lille, France (i.mitrouchev@ieseg.fr)

IÉSEG School of Management Lille Catholic University 3, rue de la Digue F-59000 Lille Tel: 33(0)3 20 54 58 92 www.ieseg.fr

Inferring Welfare from Observed Choices: An Axiomatic Approach

Guilhem Lecouteux[§] and Ivan Mitrouchev[†]

[§]Université Côte d'Azur, CNRS, GREDEG UMR 7321, 250 rue Albert Einstein, 06560 Valbonne, France.
E-mail: guilhem.lecouteux@univ-cotedazur.fr
[†]IESEG School of Management, Univ. Lille, CNRS, UMR 9221 - LEM - Lille Economie Management,
F-59000 Lille, France. E-mail: i.mitrouchev@ieseg.fr

September 2023

Abstract

Welfare economics lacks a consensus on how to infer welfare from inconsistent choices. We argue that the different approaches proposed in the literature rely on a set of values endorsed by welfare economists, defined as axioms about the structure of normative preferences and their relation to individual choices. We identify four main axioms: (i) normative individualism, (ii) choice context-independence, (iii) normative context-independence, and (iv) consumer sovereignty, which are satisfied in standard welfare economics. These axioms however become potentially incompatible when preferences are context-dependent. We show that focusing on the principles which guide welfare economists to elicit welfare from inconsistent choices open promising perspectives of research at the intersection of behavioural welfare economics and social choice theory.

Keywords. *welfare – choice – preference – context – values – social choice theory*

JEL codes. B41, D71, D90, I31

Statements and declarations. Guilhem Lecouteux acknowledges the financial support of the Mildeca through the Call for projects for Research in Public Health 2020 conducted by the Institut pour la Recherche en Santé Publique (IReSP), grant number IRESP-RSP2020-23098. Ivan Mitrouchev is supported by a grant from the Métropole Européenne de Lille (MEL). There are no competing interests.

Data availability statement. Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Acknowledgements. We thank Franz Dietrich, Marc Fleurbaey, Glen Harrison, Don Ross and Bele Wollesen for their valuable remarks, as well as the participants of the 2023 International Network for Economic Method conference in Venice for their comments. All mistakes remain ours.

1 Introduction

Standard welfare economics is based on two fundamental premises. First, it is assumed that individual choices reveal rational preferences, in the sense of a complete, reflexive and transitive relation over the set of alternatives (Varian 1987 [2014: 35]; Mas-Colell et al. 1995: 6). Second, it is assumed that the relevant normative criterion is the satisfaction of individuals' preferences, as revealed by their choices (Varian 1987 [2014: Ch. 34]; Mas-Colell et al. 1995: Ch. 16, 21). Exhibiting rational preferences allows the theorist to represent the choice of an individual as the maximisation of a utility function, which is interpreted as the individual's welfare function.[1][2] However, evidence from behavioural economics challenges the first premise, which raises the question of how to define individual welfare out of preferences which are not necessarily rational.[3] The possible discrepancy between welfare and revealed preferences is often studied by considering various notions of 'frames', defined as welfare-irrelevant features of the choice situation that can influence individual choice.[4] Although the literature is consequent and still growing, there is currently no consensus about how to infer welfare from possibly inconsistent choices.

The aim of this paper is to make the *values* that theorists endorse when providing welfare evaluation from possibly inconsistent choices explicit, and to study some implications of endorsing those values. We identify a set of four axioms, which characterises the relationship between the individual's choice and welfare: (i) normative individualism, (ii) choice context-independence, (iii) normative context-independence, and (iv) consumer sovereignty. We argue that these four axioms are satisfied in standard welfare economics and that they give a single characterisation of the individual's welfare function. However, behavioural economics challenges the validity of the choice context-independence axiom. The main implication of rejecting this axiom is that the characterisation of the individual's welfare depends on which axiom(s) is decided to be maintained. As long as choices are context-independent, the normative justification of the preference satisfaction criterion (whether it be normative individualism, normative context-independence, or consumer sovereignty) does not matter. In this case, the same welfare function is inferred from the other axioms. But when choices turn out to be context-dependent, different approaches have been suggested, which we review in Section 3: (i) behavioural welfare economics (Bernheim and Rangel 2007, 2009), (ii) behavioural paternalism (Thaler and Sunstein 2003, 2009), (iii) quantitative intentional stance (Harrison and Ross 2018), (iv) opportunity (Sugden 2004, 2018a), and (v) experienced utility (Kahneman, Wakker, and Sarin 1997). The main argument we bring

[1] By 'theorist' (she), we refer to the person – an economist, philosopher, expert, or policymaker – who is modelling the preferences of an 'individual' (he), and who may offer a normative judgement on the choice situation.

[2] In choice under risk, 'utility' is traditionally used to designate the Von Neumann Morgenstern utility of outcomes, and the utility of a prospect is characterised as the subjective expected Von Neumann Morgenstern utility of the outcomes of the prospect. Our discussion in the paper is primarily about preferences and utility defined over alternatives and not over outcomes.

[3] See McQuillin and Sugden (2012) and Chetty (2015) for overviews from different perspectives of this challenge. For a survey of empirical deviations from the standard model of rational choice (e.g. framing, intransitivity, preference reversals), see DellaVigna (2009).

[4] See Bernheim and Rangel (2007, 2009), Dalton and Ghosal (2011, 2012), Salant and Rubinstein (2008), Chambers and Hayashi (2012), Rubinstein and Salant (2012), Manzini and Mariotti (2014), among others. For a discussion, see Bernheim (2016) and Thoma (2021).

about is that finding an appropriate strategy to infer welfare from observed choices crucially depends on which values are considered to be important to conduct welfare analysis. Studying the compatibility of values is far from being unknown in the welfare literature, as it constitutes the core of social choice theory – with e.g. Arrow's (1951 [2012]) influential impossibility theorem. In this matter, our contribution can be seen as the beginning of a promising avenue of research that imports the tools of social choice theory to behavioural welfare analysis.

The rest of the article is organised as follows. We first define a formal framework that characterises how (what we refer to as) 'normative' preferences can be derived from observed choices. This framework is based on the four axioms mentioned above (Section 2). In the light of these axioms, we review the main approaches developed in the literature to infer welfare from observed choices (Section 3). We discuss the limits of each approach by highlighting the respective axioms they endorse and/or reject. We eventually propose several perspectives of research regarding the role of values in behavioural welfare analysis from a social choice perspective (Section 4). Section 5 concludes.

2 Framework

2.1 Context of Choice

We use the general notion of *context* to describe a welfare-irrelevant feature of the choice situation that can influence individual choice, in line with most theoretical models that includes framing in welfare analysis (see e.g. Bernheim and Rangel 2007, 2009). This is meant to encompass *all* kinds of factors, e.g. the order of the alternatives, the inclusion of an apparently irrelevant alternative, the mood of the moment, the weather, the time at which the choice is being made, etc.[5] Consider an individual I, who must choose an alternative x among the non-empty set of available alternatives X. Each alternative is described by a list of properties P, with $\mathcal{P}$ the set of properties. Formally, each property $P \in \mathcal{P}$ is a function assigning to each alternative $x \in X$ a value $P(x)$ from some range. In the case of a binary property, the range is $\{0; 1\}$, where $P(x) = 1$ means that x has the property and $P(x) = 0$ means that x does not have the property. More generally, the range could be some interval of values, where $P(x)$ represents the degree to which x has the property – e.g. the distance between the alternative x and a reference point. Properties can either refer to intrinsic properties of the alternatives (e.g. colour, shape) or extrinsic properties of the alternative (e.g. social norms).

We consider different types of properties: (i) motivational properties $P \in \mathcal{M}_I \subseteq \mathcal{P}$, (ii) known properties $P \in \mathcal{K}_I \subseteq \mathcal{P}$, and (iii) relevant properties $P \in \mathcal{R}_I \subseteq \mathcal{P}$. Before going further, it is important to stress here that the sets $\mathcal{M}_I$, $\mathcal{K}_I$ and $\mathcal{R}_I$ are the *theorist's* representation of the choice problem faced by I (meaning that nothing guarantees that the individual would agree with the theorist's representation). Motivational properties are the properties which influence the actual choice of the individual, known properties are the properties of which the individual is aware – i.e. when considering

[5] Our definition of context is therefore extremely general and does not refer to the violation of a particular axiom of rational choice, such as independence of irrelevant alternatives (Tversky and Simonson 1993).

the alternatives, the individual can determine the value $P(x)$ – and relevant properties are the properties which are normatively-relevant for the individual – i.e. the properties that determine whether an alternative is 'better' than another for the individual. The set of motivational, known, and relevant properties may overlap, and there is *a priori* no relation of inclusiveness between $\mathcal{M}_I$, $\mathcal{K}_I$, and $\mathcal{R}_I$.

As an example, imagine an election where I is voting and politician Smith is one of the candidates. Smith is bold, promotes a centrist political agenda, and also sets up a team of supporters who artificially increase his visibility on social media. We have here several properties characterising Smith, which could be represented as follows.

- $P_b(Smith) = 1$, meaning that the property 'boldness' is satisfied.
- $P_p(Smith) = 0.5$, meaning his political agenda, on a range of real numbers from 0 to 1 – representing whether he is on the left or right side of the political spectrum – is in the middle.
- $P_v(Smith) = 80$, giving a score of visibility on social media, from e.g. 0 to 100.
- $P_m(Smith) = 1$, meaning the property 'manipulation' is satisfied.

Suppose that $\mathcal{K}_I = \{P_b, P_p\}$, $\mathcal{R}_I = \{P_p, P_m\}$, and $\mathcal{M}_I = \{P_p, P_v\}$. The voter is aware of Smith's political agenda and of his boldness, while he considers that only his political agenda is relevant for his vote. However, he does not know that Smith is a manipulator, while this should – at least from the perspective of the theorist – also be relevant for his vote (Smith being not necessarily trustworthy). Furthermore, he does not know that social media visibility – which is not relevant for his vote – may however influence his actual vote. We have here a situation in which a property is relevant, motivational, and is known (Smith's political agenda), another which is also relevant, but neither motivational nor known (Smith's manipulation), a property which is motivational, but neither known nor relevant (Smith's visibility), and another which is known, but neither relevant nor motivational (Smith's boldness).[6]

Our definition of the context is based on the premise that it refers to what *we* theorists consider as the 'irrelevant' properties of the choice problem (Bacharach 2006: 13). In particular, the set of relevant properties is the theorist's own representation of the choice problem at stake – although we cannot be *a priori* certain that the individual himself considers (or would consider, upon careful scrutiny) these properties as being relevant.[7] For simplicity, we assume that the theorist correctly identifies the set $\mathcal{M}_I$, i.e. she precisely knows the properties that influence the choice of the individual.[8] Formally, a *context property* is a property that is motivational but not relevant: $P \in \mathcal{C}_I = \mathcal{M}_I \setminus \mathcal{R}_I$.

[6]We could have completed this illustration with other cases, e.g. motivational and known, but not relevant properties, such as the weather on polling day, which may lead the voter to abstain. The main point is that we impose no constraint on the relationship between the three sets.

[7]We remain silent on the adequate *perspective* from which the relevant properties and individual welfare should be evaluated, which could either be the current individual's judgement, his counterfactual enlightened judgement as estimated by the theorist, or the individual's ability to aggregate different judgements taken from different perspectives. We explore this question in a companion paper [anonymised, *forthcoming*].

[8]Relaxing this assumption would lead us to consider that the theorist could have a wrong representation of the choice problem, which is a complication we prefer to avoid.

A context is any combination $\gamma = (\gamma_P)_{P \in \mathcal{C}_\mathcal{I}} \in \Gamma$ of values of the context properties. In the example above, there is only one property – visibility on social media – that is motivational and not relevant, i.e. $\mathcal{C}_I = \{P_v\}$, and the context is defined as the set of scores of visibility on social media of the different candidates.

2.2 Choice and Welfare

Given our definition of motivational properties, individual choice is a function that maps each subset of motivational properties $\mathcal{M}_I$ to a choice function over menus of alternatives from X.[9] This model bears some similarities with Dietrich and List's (2013a, 2013b) model of 'motivationally salient properties' and their approach to model context-dependent preferences (Dietrich and List 2016). Knowing that a context property is motivational by definition, we define I's choice as a function of the context γ, and denote it $C_\gamma \subset X \times X$. We interpret C_γ as a choice ranking: '$x\ C_\gamma\ y$' reads as 'I chooses x over y in context γ'. It means that, when asked to choose between x and y in a context γ, I chooses x. We do not make any assumption about the properties of C_γ, e.g. whether it is transitive or not, or whether it could be interpreted as desires or motives for actions. Instead, we consider it as an analytical index aimed at representing the behaviour of the individual.

We define $\succ_\gamma \subset X \times X$ as the normative preference of the individual in context γ, which is the ranking that characterises the individual's welfare.[10] While C_γ represents the actual choice of the individual in context γ, $\succ_\gamma$ represents the preference that he ought to satisfy in order to maximise his welfare. The distinction between C_γ and $\succ_\gamma$ allows us to distinguish between the 'descriptive' and the 'normative' aspects of individual decision-making. For convenience, assume that C_γ and $\succ_\gamma$ are complete relations, $\forall \gamma \in \Gamma$.

While we can directly observe individuals' choices, this is not true of their normative preferences. Given our definition of motivational and relevant properties, an intuitive approach would be to define the normative preferences of an individual as the preferences he would reveal if he was only motivated by relevant properties, i.e. $\mathcal{M}_I = \mathcal{R}_I$. This is the strategy of standard welfare economics, which defines normative preferences $\succ$ as the preferences revealed by the individual's choice. However, the challenge raised by behavioural economics is that there may exist properties which are motivational but not relevant, and that $\mathcal{R}_I$ is the theorist's prior belief about what *she* thinks matters for the individual (e.g. that Smith is a manipulator).

As an illustration, consider the Asian disease experiment of Tversky and Kahneman (1981: 453). An unusual Asian disease is expected to kill 600 individuals. Subjects were

[9] A menu is a non-empty set $Y \subseteq X$ of feasible alternatives, and a choice function maps each menu Y from some set of possible menus to an alternative in Y, representing the alternative chosen from this menu. We say 'some set of possible menus' rather than 'all menus', because many combinations of alternatives (such as the totality of X) do not define a possible menu, as the alternatives have mutually inconsistent properties.

[10] Different terminologies can be found in the literature: true, authentic, laundered, implicit preferences (among possibly others). Our concept of normative preference is general, meaning that we do not say something particular on the *kind* of preference that actually makes the individual better off, such as a preference that would satisfy some norms of rational choice.

asked to choose between two different health programs: a certain and a risky alternative. The choice between the two programs can be framed in terms of gains or losses. The % below corresponds to the share of subjects who choose the program in Tversky and Kahneman's experiment, and N corresponds to the total number of subjects per frame.

Frame 'gain' [N = 152]
A: 200 people will be saved [72%]
B: 1/3 probability that 600 people will be saved,
and 2/3 probability that no people will be saved [28%]

Frame 'loss' [N = 155]
C: 400 people will die [22%]
D: 1/3 probability that nobody will die,
and 2/3 probability that 600 people will die [78%][11]

This experiment suggests that the framing (in terms of gains *vs.* losses) is a motivational property, although we (as theorists) can reasonably doubt whether it is a relevant property of the choice problem. From a purely consequentialist perspective, the two alternatives are indeed identical. In this type of situation, with a clear influence of a context property, it may be more difficult to identify individuals' normative preferences.

2.3 Values in Welfare Analysis

One way to clarify the debate around the challenge of inferring welfare from observed choices is to identify the *values* endorsed by theorists. We propose to represent such possible values through a set of axioms that relates choice to welfare, namely: (i) normative individualism, (ii) choice context-independence, (iii) normative context-independence, and (iv) consumer sovereignty. We can show that axioms (i) and (ii), when combined, imply axioms (iii) and (iv), which offer an unambiguous way to define normative preferences. The challenge raised by behavioural economics is however that axiom (ii) does not hold in many situations, which means that axioms (iii) and (iv) must be *postulated* in order to derive normative preferences. An additional difficulty is that the characterisation of normative preferences that is derived from axiom (iii) is not anymore compatible with the characterisation derived from axiom (iv). That is, the welfare function that would be inferred by maintaining normative context-independence is different from the one that would be inferred by maintaining consumer sovereignty, meaning the theorist must choose one of these two principles before eliciting normative preferences.

We now formulate and comment these axioms in turn. According to the principle of normative individualism, the proper locus of normative concern is individual persons, whose values and situations should be taken into account when debating ethical issues such as policy or justice.[12] We translate this principle in our framework as follows.

[11]This experiment is a survey response based on an unincentivised hypothetical choice task. For the experimentalist reader who would prefer to have an example with an incentivised choice task, we refer him/her to known discrepancies between different preference elicitation methods, such as the discrepancy between certainty equivalence and probability equivalence (Hershey and Schoemaker 1985).

[12]See Ross (2005: 220-222) for a contemporary definition. This principle has obviously deeper ideological and philosophical roots, that could be found in foundational references such as J. S. Mill's *Principles of Political Economy* and *On Liberty*.

AXIOM 1. Normative Individualism (NI). For any pair of distinct alternatives (x, y) and context $\gamma \in \Gamma$, $\succ_\gamma$ must be such that:

i. $x \succ_\gamma y$ only if there exists at least one context γ' such that $x\ C_{\gamma'}\ y$

ii. $x \succ_\gamma y$ if $x\ C_{\gamma'}\ y$, $\forall \gamma' \in \Gamma$

This principle establishes a close relation between the choice and the normative preference of the individual. x can be considered as better than y in context γ only if there exists at least a context γ' in which he would indeed choose x (condition i.). In other words, x cannot be better than y if the individual never chooses x over y. Furthermore, if the individual always chooses x independently of the context, then the x is necessarily better than y (condition ii.). The fundamental idea of this definition is that individual welfare should not be set *a priori* but rather inferred from actual choices, although possibly – but not necessarily – in a different context from the current one. If there does not exist any context in which I would choose x, then x cannot be better than y. And if I always chooses x, then x must be better than y. Since the two conditions are not complementary, the principle of NI remains silent on cases where the choice between x and y depend on the context. The second axiom we propose is *choice context-independence*, and offers a way to address the indetermination of NI:

AXIOM 2. Choice Context-Independence (CCI). $\forall \gamma, \gamma' \in \Gamma, C_\gamma = C_{\gamma'}$

CCI states that I's choice does not depend on the context in which he is embedded. According to our framework, it means that the set of context properties is empty: any motivational property is necessarily relevant, and *vice-versa*. Unlike NI, which is a normative principle, CCI is an informal assumption about how individuals actually behave in a choice situation. A recurrent finding in behavioural economics is, however, that choices seem to be context-dependent, with e.g. framing effects leading to violation of the invariance principle, such as in the Asian disease experiment.[13] We now turn to the normative counterpart of CCI:

AXIOM 3. Normative Context-Independence (NCI). $\forall \gamma, \gamma' \in \Gamma, \succ_\gamma = \succ_{\gamma'}$

NCI means that the normative preferences of the individual do not depend on the context of choice, i.e. there exists a stable (context-independent) preference relation that determines the individual's welfare. This principle has some normative appeal – at least from the theorist's perspective – since it means that the individual's welfare only depends on what the theorist thinks is relevant for the individual. Once the theorist has identified a set of relevant properties, NCI guarantees that we can define a welfare function. If this was not the case, the welfare associated to a given alternative could vary depending on the context of choice, resulting in a welfare function which is unstable across contexts. The fourth axiom we propose is the principle of consumer sovereignty:

[13]The concept of frame, as initially introduced, refers to the violation of the invariance principle, according to which the preference order between alternatives should not depend on the manner in which they are described (Tversky and Kahneman 1986).

AXIOM 4. Consumer Sovereignty (CS). $\forall \gamma \in \Gamma, C_\gamma = \succ_\gamma$

CS embodies the idea that the individual himself (and nobody else) is the best judge of what makes him better off.[14] More specifically, the principle states that the normative preferences of an individual over X precisely correspond to his choices over X. To put it differently, any motivational property is necessarily relevant. This means that the set of contexts is empty because the theorist prefers to 'extend' the set of relevant properties to include all the properties that influence the individual's choice.

We can note several conditions of inclusion and compatibility between the four axioms. First, CS is more restrictive than NI. By construction, CS respects the condition i. of NI, although it imposes that $\succ_\gamma$ necessarily corresponds to C_γ (while according to NI, $\succ_\gamma$ is known for sure only if the choice between two alternatives remain the same across contexts). Second, NCI and CS are often incompatible. If we have $\gamma, \gamma' \in \Gamma$ such that $C_\gamma \neq C_{\gamma'}$ (i.e. choices are context-dependent, and CCI is rejected) then CS implies $\succ_\gamma \neq \succ_{\gamma'}$, which violates NCI. Third, NI and NCI can be compatible (although not necessarily), as long as for all $x, y \in X$, if $x \succ_\gamma y$, we can find $\gamma' \in \Gamma$ such that $x\ C_{\gamma'}\ y$. Lastly, the combination of CCI and NI imply both NCI and CS (proofs in appendices A.1 and A.2).

NI embodies the idea that normative preferences must be derived from observed choices, which is a constitutive principle in standard welfare economics. Furthermore, since there is no reference to a notion of 'context' in standard welfare economics, it is as if CCI is assumed (choices do not depend on the context). CS therefore holds in standard welfare economics, as well as NCI. Observed choices necessarily reveal the underlying normative preferences of the individual, which directly allows the theorist to elicit them. However, if choices can be context-dependent, then we have to reject CCI, and CS and NCI are not compatible anymore. The challenge raised by behavioural economics is therefore to derive normative preferences when CCI does not hold. NI remains silent on cases where choices depend on the context. Consequently, it seems that either NCI or CS has to be postulated if we want to derive the normative preferences of the individual.

3 Literature Review

In the light of the four axioms formulated above, we discuss in detail the main alternatives that have been suggested in the literature to derive normative preferences when CCI does not hold. We categorise the literature as follows: (i) behavioural welfare economics, (ii) behavioural paternalism, (iii) quantitative intentional stance, (iv) opportunity, and (v) experienced utility.

[14] This concept has been formulated by Hutt in his *Economists and the Public*, and formulated again in an exchange with Fraser as 'the controlling power exercised by free individuals, in choosing between ends, over the custodians of the community's resources, when the resources by which those ends can be served are scarce' (Hutt 1940: 66). While the concept originally referred to the means-end relation in consumer behaviour (in the spirit of Robbins' definition of economics), it later and dominantly referred to the principle that 'arrange[s] for everybody to have what he prefers whenever this does not involve any extra sacrifice for anybody else' (Lerner 1972: 258).

3.1 Choice-Based Framework

The choice-based framework (Bernheim and Rangel 2007, 2009) is to extend standard choice welfare analysis to situations where individuals make 'anomalous' choices of various types commonly identified in behavioural economics. In this approach, frames are, by assumption, irrelevant to the definition of individual welfare. Frames are akin to the context properties in our framework, that are motivational but not relevant. The main principle of this approach is to conduct welfare analysis by identifying the operational misunderstandings of the relationship between means and outcomes (which are treaten as 'mistakes') that can be elicited with the use of cognitive data (Bernheim 2016). The process consists in tracking context properties by identifying inconsistent choices, and then to make normative evaluation only on the sets of choices for which we cannot reasonably identify the influence of a context property. The individual welfare function is then derived from this restricted set of choices.

In this approach, the strategy is to 'rescue' CCI. It is well recognised that individuals' preferences may change across contexts. However, for the sake of welfare analysis, CCI is maintained by restricting the choice domain that serves as the input in welfare analysis to 'non-ambiguous' choices. This approach may be considered as a pragmatic strategy to the challenge of inferring welfare from observed choices. In this respect, it extends the revealed preference framework by taking into account the cognitive processes of individuals without modifying its overall principle, according to which x is unambiguously preferred to y if and only if y is never chosen when x is available. NI is therefore preserved. As CCI is maintained by construction of the set of choices under consideration, NCI and CS are also maintained in the restricted set of choice data that is considered to be 'unbiased'. Removing the 'ambiguous' data from welfare analysis implies, however, that the theorist cannot make normative evaluation in cases where individual choice is 'too' inconsistent. This means that the range of situations which can be studied is rather restricted, and the theorist cannot conduct welfare analysis in situations where choices highly vary across contexts.

3.2 Behavioural Paternalism

Behavioural paternalism characterises individual welfare as the satisfaction of preferences that are not distorted by cognitive biases.[15] A possible interpretation of this literature is that an individual would make 'adequate' choices in a context-free situation, i.e. without cognitive limitations. Translated to our framework, CCI is here explicitly rejected while NI is maintained.[16] Here the rejection of CCI leads to the rejection of CS (since it is considered that individuals can make mistakes), while NCI is maintained (the adequate context to infer normative preferences is when the individual is not influenced

[15]The most influential account is given by Thaler and Sunstein (2003, 2009) in their defence of libertarian paternalism and in their popular *nudge* approach. Similar forms of paternalism have been advocated in Camerer et al. (2003) (asymmetric paternalism), Loewenstein and Ubel (2008) (light paternalism), and Dalton and Ghosal (2011) (soft paternalism). We label these approaches under the general term of 'behavioural paternalism', where the theorist aims at enhancing the welfare of boundedly rational individual with no (or minor) cost to rational individuals.

[16]In this literature, the NI principle refers to the 'as judged by themselves' clause (Thaler and Sunstein 2009). See Sunstein (2018) and Sugden (2018b) for a debate about the meaning and possibility to satisfy this clause.

by context properties).

Within our framework, we see two difficulties for behavioural paternalism. First, nothing guarantees that the individual's inner rational agent – i.e. the counterfactual individual who is free from cognitive limitations – would reveal context-independent preferences, as argued by Infante, Lecouteux, and Sugden (2016). To put it differently, even if the set of motivational properties is restricted to the set of relevant properties, nothing guarantees that the individual will make context-independent choices. Indeed, choices derived from relevant properties may not necessarily be complete, in which case using the context to choose between two alternatives may be considered as an acceptable choice rule for the individual. In this case, normative preferences would be considered as context-dependent as well, which eventually leads to violate NCI. As a result, it may not be possible to define a stable (context-independent) welfare relation from individuals 'de-biased' preferences.

Second, it is not obvious that the theorist can correctly identify the context properties, which are motivational but not relevant.[17] Behavioural paternalism presupposes that the set of relevant properties $\mathcal{R}$, as represented by the theorist, precisely corresponds to the properties that are relevant to the individual. This is a more general issue related to the disentanglement among motivational properties of the sets of relevant and context properties. Even if $\mathcal{M}$ is correctly identified, the theorist cannot know *a priori* whether a motivational property is relevant or not. Let us take the example of Smith's election. The theorist considers that the fact that Smith manipulates social media is relevant (because it reveals he is not trustworthy), while the individual could perfectly be fine with it – e.g. he considers it is part of an acceptable electoral strategy, and therefore that being a manipulator is not relevant for his final choice. Similarly, in the Asian disease experiment discussed earlier, the theorist cannot know *a priori* whether the individual ought to be risk-averse or risk-seeking. This suggests that NI may not hold in behavioural paternalism, despite the narrative promoted by tenants of this literature. Indeed, behavioural paternalism imposes *consistency* across contexts as a normative criterion, which appears to be more controversial than usually considered, and would require additional justification.[18]

3.3 Quantitative Intentional Stance

Another approach that intends to make welfare inferences from observed choices while acknowledging that CCI is invalidated is the *quantitative intentional stance* proposed by Harrison and Ross (2018, 2023). This approach is based on Dennett's (1987) externalist account of preferences and beliefs. Those are not defined as inner mental states that are the cause of individual behaviour, but rather as attributions to oneself and others that make one's behaviour socially understandable. In this approach, looking for a notion of welfare does not require investigating individuals' mental states. It requires

[17] See Rizzo and Whitman (2009) who refer to this problem as the 'knowledge problem' in behavioural paternalism. Note that such a problem is far from being unknown in public economics, where a fundamental task of the theorist is to set up an incentivised mechanism so that individuals reveal their 'true' preferences (Atkinson and Stiglitz 2015: Ch 16.6). In this framework, the problem is however rather of *trustworthiness* between the theorist and individuals than of welfare elicitation *per se*.

[18] See Arkes, Gigerenzer, and Hertwig (2016) and Lecouteux (2021) for an extensive analysis of the lack of normative justification of consistency.

interpreting individual behaviour in terms of the theorist's own language of subjective expected utility. As an illustration, Harrison and Ng (2016, 2018) and Harrison and Ross (2018) characterise the risk preferences of individuals by eliciting the most likely preference structure (expected utility or rank-dependent expected utility) in simple experimental tasks, and then use those risk preferences as the welfare metric for choices among insurance products or portfolios. The articulation between the lab and the field is crucial in this approach, since the lab is the adequate environment from which the theorist can infer her prior beliefs about the risk preferences and beliefs of the individual.[19] The elicitation in the lab of the theorist's prior beliefs about the welfare of the individuals also allows her to anticipate the welfare effects of any intervention in the field (Harrison, Morsink, and Schneider 2020), while most typical nudge interventions merely postulate *a priori* the welfare of the individual.

According to our framework, the quantitative intentional stance rejects CCI and keeps NI, as well as NCI. The suggestion according to which welfare can be measured in lab experiments is justified by considering that there is a lower risk of context-dependence in the lab, which offers an environment where the theorist can reasonably assume that the only properties considered by the individual are relevant. In this sense, it offers an operational measure to determine the normative preferences (or at least, the welfare distribution) of individuals. In this approach, normative preferences correspond to the actual choices individuals would exhibit in a lab experiment, where the 'noise' and uncertainty of the surrounding environment is minimised. The relative arbitrariness of the definition of welfare, as the most likely (econometrically speaking) utility structure characterising the individual preferences and beliefs, is here explicitly recognised as the theorist's prior. There is therefore a possibility of 'mistake' (Harrison and Ross 2023: Chap. 2.E), and CS is rejected – even though their definition in terms of structural models of noisy decision-making is much more precise than the almost pathological description found in behavioural paternalism with individuals afflicted by many biases (Lecouteux 2023). Furthermore, from a more pragmatic perspective, the theorist in this approach is not an abstract social planner but a hired consultant advising an actual client (e.g. a person employed by a bank who aims to improve the financial choices of his clients). This means that even if CS is rejected, it is made with the explicit consent of the client, who expresses his willingness to delegate his states of affairs to the theorist. The quantitative intentional stance – compared to the choice-based framework discussed previously – offers an operational approach to welfare analysis, but still faces a restriction: it is only applicable to 'preferences that violate [expected utility theory] but [which] are nevertheless well ordered' (Harrison and Ross 2018: 22).

3.4 Opportunity

Sugden (2004, 2018a) proposes a distinctive approach in this debate, by rejecting NCI and shifting the normative focus from welfare to opportunity. This strategy values indi-

[19] This is because such experiments are considered as 'small worlds' – in Savage's (1954) terms – where subjective expected utility can hold. Practically speaking, the strategy consists in estimating, from a set of choices between risky lotteries, the distribution of risk preferences and subjective beliefs of the individual, rather than a single characterisation (e.g. taking the mean to estimate the parameters) of the risk preferences and beliefs (Gao, Harrison, and Tchernis 2023). Unlike the other approaches, the quantitative intentional stance is primarily developed to analyse situations of choice under risk, with the elicitation of (von Neumann-Morgenstern) utility functions and subjective beliefs.

vidual *freedom* of choice rather than their actual choices. The role of the theorist is not to make policy recommendations that maximise individual welfare, but to ensure that institutions are designed in a way that it is in the interest of each individual to accept the rule of those institutions. A typical example of such an institution is the market, which maximises the opportunity sets of market participants and thus facilitate the realisation of mutual benefit – in which case the market is rather seen as a cooperative than a competitive institution (Sugden 2018a). Unlike the rest of the literature discussed in the present article, the theorist has no role in identifying the relevant properties of a choice situation, as she does not aim at making normative evaluation from individuals' preferences at all.[20] The individual I is seen as 'a continuing locus of responsibility', treating his past, present and future actions as his own, whether or not these actions were or will be what he would like them to be now (Sugden 2004: 1018). Such a quality of 'responsible person' gives normative authority to the judgement of the individual on his own actions. That is, it is up to individuals to choose as they prefer, even though their choices are likely to be context-dependent, and therefore highly inconsistent. Translated to our framework, this approach rejects CCI and NCI, and the adequate context for the definition of normative preferences simply corresponds to the *current* context of a choice. CS is maintained and gives a direct way to define normative preferences.

The opportunity approach imposes a strong version of NI, where all contexts *must* be considered as relevant for individual welfare. Yet it remains silent on cases that may appear relatively concerning, such as (i) self-acknowledged failures of self-control (e.g. drug addiction) and perhaps most importantly, (ii) cases where individuals' preferences are strongly influenced by unknown properties (e.g. aggressive marketing or adaptive preferences), whose knowledge may result in changing their choice. One example of restriction of the opportunity approach is that it may be difficult to disentangle cases of adroit marketing (such as a baker who prominently displays her nicest desserts rather than offering them already wrapped in cellophane) and cases of manipulative techniques such as using ambient scent in supermarkets as a strategy to induce different moods and desires (Akerlof and Shiller 2015). In this approach, there is no decisive criterion to identify which cases can be considered or not as outright forms of fraud and deception on behalf of firms, which could result in violating the rules of fair competition – that each individual is initially expected to accept.

3.5 Experienced Utility

Lastly, some authors propose to reject *choice* as the relevant criterion and to rely on the hedonic quality of an experience, captured by the concept of *experienced utility*. In contrast with decision utility, which refers to the weight given to an outcome in a decision (and which is therefore based on individuals' choices), experienced utility refers to the actual experience in choosing an alternative over another, in the sense of the Benthamite pain/pleasure dichotomy (Kahneman, Wakker, and Sarin 1997). Translated into our framework, experienced utility is derived from the satisfaction of normative preferences. In this approach, it is explicitly acknowledged that decision utility is context-dependent, and therefore that CCI is rejected. Unlike in the choice-based framework and behavioural paternalism, NI is however rejected. This is because experienced

[20] See Mitrouchev (2019) for a detailed assessment of this approach compared to behavioural paternalism.

utility is a criterion that evaluates the outcome of a decision independently from (or external to) the individual system of preferences. In this matter, normative preferences are not defined from individual choices but are postulated *a priori*. Specifically, the approach suggested by Kahneman (1999) is to define 'objective happiness' according to a set of normative rules that are external to the subject. The experienced utility approach therefore keeps NCI. Since normative preferences are defined independently of individual behaviour, it also rejects NI, and therefore CS.

One major restriction we see with this approach is that rejecting NI *and* CS leads to a full delegation of individual welfare to the theorist. The obvious difficulty of such an *a priori* account is its arbitrary definition of welfare, and its ignorance of the individual agency and autonomy.[21] Kahneman (1999), for instance, argues that 'policies that improve the frequencies of good experiences and reduce the incidences of bad ones should be pursued *even if people do not describe themselves as happier or more satisfied*' (15, our emphasis), which may raise significant ethical issues.

4 Social Choice and Behavioural Welfare Analysis

4.1 Values and Behavioural Welfare Analysis: A Summary

Table 1 below summarises the positions of the approaches we reviewed in Section 3. A checkmark means that the axiom is maintained. A crossmark means that the axiom is rejected.

Table 1: Axiom check for each normative approach

	NI	CCI	NCI	CS
Choice-based framework	✓	✓	✓	✓
Behavioural paternalism	✓	✗	✓	✗
Quantitative intentional stance	✓	✗	✓	✗
Opportunity	✓	✗	✗	✓
Experienced utility	✗	✗	✓	✗

Based on our analysis, the literature suggests that rejecting CCI implies either to maintain NI and NCI and reject CS (behavioural paternalism and quantitative intentional stance), or maintain NI and CS and reject NCI (opportunity) or maintain NCI alone while rejecting NI and CS (experienced utility). The fundamental problem that appears is a choice between rejecting CS – which means that the theorist considers

[21] About individual agency, we can, for example, refer to Nozick's (1974) 'pleasure machine' thought experiment. The thought experiment consists in asking whether we would prefer to be connected to a machine that would maximise our happiness rather than living the real life. Nozick provides three arguments why it is not desirable to do so. First, we want to *do* certain things, not just have the experience of doing them. Second, (in relation to the first point), this is because we want to be a certain kind of person and not 'an indeterminate blob floating in a tank' (43). Third, plugging into an experience machine limits us to man-made reality, where there is no contact with a 'deep reality'. About individual autonomy, it can be noted that, even though the other approaches discussed above do not all explicitly engage with a notion of autonomy, the idea remains implicit in their definition of normative preferences and whether behavioural economics challenges consumer sovereignty (Lecouteux 2022a).

the possibility that individuals make errors – and rejecting NCI – which means that the theorist is not able to define a stable welfare function anymore. Maintaining both CS and NCI requires maintaining CCI, which means remaining ambiguous on many possible policy-relevant cases.

One way to frame the choice between NCI and CS is to refer to the debate between the aggregative welfare approach to normative economics and the non-aggregative approach found in social contract theory. This similarity between behavioural welfare analysis and social choice theory is mentioned by Sugden (2018a) in the preface of the *Community of Advantage* (viii-ix), where he draws a parallel between his critique of Sen's impossibility of a Paretian liberal (Sen 1970; Sugden 1985) and his proposition of the individual opportunity criterion (Sugden 2004). We see here potential bridges between behavioural welfare analysis and social choice theory, in particular regarding the three following questions:

i. What are desirable values regarding the definition of normative preferences?

ii. Are those values compatible, when considering the intrapersonal aggregation of individual preferences?

iii. What would be the outcome of an intrapersonal bargaining process between conflicting preferences?

We briefly sketch below some lines for further research that would address these three questions.

4.2 Preference Integration

We propose that NI should constitute the basis of behavioural welfare analysis, i.e. that welfare evaluation should ultimately depend on the individual's choices – even though it may recognise the possibility of errors (i.e. there may exist $\gamma \in \Gamma$ for which $C_\gamma \neq \succ_\gamma$). This rejects approaches based on 'experienced utility', which rely on an arbitrary definition of welfare, and which may raise some ethical problems. As discussed above, we think that CS may be a too strong formulation of NI, since all motivational properties must be treated as relevant, although there may be some doubts in various situations (e.g. self-acknowledged failures of self-control or addictions).[22]

Since NI remains silent on cases for which choice is context-dependent, we need to identify general principles (values) to identify desirable properties of normative preferences – and how they relate to choice across different contexts. The second condition of NI is a typical illustration: $x \succ_\gamma y$ if $x\, C_{\gamma'}\, y$, $\forall \gamma' \in \Gamma$, which relates to a condition of unanimity in social choice theory such as, if x is preferred to y by all individuals, then x must be socially preferred. Our proposition is that many paradoxes or impossibility theorems known in social choice theory (e.g. Arrow (1951 [2012]) or Sen (1970 [2017]) can be

[22] A possible solution would be to have a criterion that could disentangle between situations of choice where the individual is 'sovereign' in his choice, and situations for which his autonomy is seriously hampered – with an appropriate definition of 'autonomy', depending on one's ontological commitment of the definition of the agent. This, however, also remains an open question. See Lecouteux (2022b).

transposed at the intrapersonal level if we treat an individual as a collection of subpersonal selves defined over contexts. As an illustration, desirable properties for normative preferences could mimic those of Arrow's (1951 [2012]) impossibility theorem.[23]

- **Unrestricted domain.** For any set $\{C_\gamma\}_{\gamma\in\Gamma}$ of a choice function, there exists a normative preference $\succ$ that is reflexive, transitive, and complete. In other terms, we should be able to define a welfare function for the individual, for any logically possible set of context-dependent preferences.
- **Unanimity (or Pareto property).** $x \succ y$ if $x\ C_\gamma\ y$, $\forall\gamma \in \Gamma$. In other terms, if an alternative is always chosen over another, it must be normatively preferred.
- **Independence of irrelevant alternatives.**[24] if $\langle C_\gamma\rangle(x;y) = \langle C^*_\gamma\rangle(x;y)$, then $\succ (x;y) = \succ^* (x;y)$, $\forall C, C^* \in X \times X$. In other terms, normative preferences between two alternatives should depend only on choices between these two alternatives.
- **Non-dictatorship.** $\nexists\gamma^* \in \Gamma$ such that, $\forall\{C_\gamma\}_{\gamma\in\Gamma}, \succ = C_{\gamma^*}$. In other terms, there is no context whose choice function systematically determines the normative preferences.

Unrestricted domain means that we can always derive a welfare relation from individual choices. This is verified with the experienced utility and behavioural paternalism approaches, but neither with the opportunity approach (since C_γ is not always transitive), the choice-based framework (which leaves ambiguous data aside), nor the quantitative intentional stance (which requires a minimal degree of regularity in the choice patterns). Unanimity is the second part of NI, and is thus found in all approaches but experienced utility. Non-dictatorship is verified in the opportunity approach, while being clearly violated in behavioural paternalism, which imposes choice in a 'context-free' situation as the legitimate one.[25]

From this brief (and incomplete) overview of the different approaches with respect to the values listed here, we can see that the approaches that maintain NCI (hence allowing the definition of a welfare function) violate at least one of the principles. From a methodological point of view, the problem of preference *integration* is closely related to the problem of preference *aggregation* in social choice theory. The main difference between preference integration and preference aggregation is that the former is concerned with *intrapersonal* aggregation of preferences – aggregating different preferences belonging to the same individual – while the latter is concerned with *interpersonal* aggregation of preferences – aggregating different preferences of distinct individuals. We can mention in the existing (relatively limited) literature addressing this point, the works of Steedman and Krause (1986) and Binder (2014), which characterise the conditions under which the aggregation is possible at the intrapersonal level. In a nutshell, they suggest that an aggregation may only be possible if the degree of conflict between the various choices of the individual is low.

[23] We drop the subscript γ for $\succ_\gamma$, since we have to respect NCI in order to define a welfare function.

[24] $\langle C_\gamma\rangle(x;y)$ denotes the ranking between x and y induced by the choice functions $\{C_\gamma\}_{\gamma\in\Gamma}$.

[25] The relationship between independence of irrelevant alternatives and the various approaches reviewed earlier is less straightforward, which is the reason we prefer not to discuss it here.

4.3 Social Contract

If we tackle the challenge of behavioural welfare analysis from the perspective of social contract theory rather than preference integration, various research questions emerge. In a companion paper [anonymised, *forthcoming*], we propose that normative evaluations should be based on the intrapersonal confrontation of different perspectives on the same choice problem, knowing that those perspectives are themselves context-dependent. This confrontation of perspectives respects NI while not taking CS *prima facie*. Normative preferences are indeed fundamentally related to individual choices, while we recognise the possibility of mistakes – i.e. choices made in certain contexts that, viewed from the perspective of another context, are not accepted by the reflexive individual. This emphasis on intrapersonal bargaining is noted by Hédoin (2015), who argues that 'behavioral economists have totally ignored the solution of Coase (1960), which consists in letting the agent's various selves to (interpersonally) bargain over the internalities' (78). If assumptions about bargaining between individuals make sense when transposed to a bargaining between selves, some results on *social* bargaining could likely be transposed to *individual* bargaining. As an illustration, since a notion of 'sub-coalition of selves' probably makes less sense than a sub-coalition of players, we can imagine that conditions for coalitional stability for the Coase theorem could be more easily met (Aivazian, Callen, and Lipnowski 1987). We can also imagine that the problem could be addressed with the tools of cooperative game theory (Gonzalez, Marciano, and Solal 2019), or with a model of intra-personal team reasoning (Gold 2021). We emphasise that the literature offers a vast variety of tools with which to address the problem of welfare evaluation from inconsistent choices. In this matter, we can formulate more procedural normative criteria on the process through which normative preferences can be formed by confronting different perspectives.

5 Conclusion

Welfare economics lacks a consensus on how to infer welfare from inconsistent choices. We argue that the different approaches proposed in the literature rely on the different values endorsed by welfare economists, that we define as axioms about the relation between observed choices and normative preferences. We build our analysis on the notion of *context* of choice, in terms of 'motivational but not relevant' properties. This allows us to clearly highlight that the distinction between context properties and relevant properties are first and foremost the theorist's representation. We identified three values (in particular) that characterise the structure of normative preferences: normative individualism, normative context-independence, and consumer sovereignty. Standard welfare economics does not consider the possibility of context property (i.e. properties of the alternatives that are motivational but not relevant). In our framework, this means that CCI is assumed. The direct consequence is that both NI and CS have the same characterisation of the individual's normative preferences. Furthermore, NCI is satisfied in this case, meaning it is possible to define a stable welfare function. The challenge raised by behavioural economics is that, without CCI, NI remains silent on the normative preferences for which individual choice is context-dependent.

We propose that NI must be maintained as the basis of welfare analysis, meaning that individual normative preferences must be related to their own choices (and not

imposed by the theorist). If we maintain CS (opportunity approach), then normative preferences are context-dependent, which means that NCI is rejected, and we cannot define a stable welfare function. Furthermore, maintaining CS without CCI implies that all motivational properties must be considered as relevant, although we may find disturbing cases (e.g. addictions and deceptive behaviours). Maintaining NCI, which is necessary if the theorist wants to offer welfare evaluations, implies rejecting CS, and recognising the possibility of errors – unless we remain explicitly agnostic about ambiguous choices (choice-based framework). The definition of welfare is then more or less arbitrary when we reject any reference to individual choice (experienced utility), or when we consider the counterfactual enlightened choices of the individual as the 'correct' preferences (behavioural paternalism), or when we calibrate the theorist's priors as the most likely utility structure of the individual in controlled experimental tasks (quantitative intentional stance).

Our main conclusion is that identifying a way to infer welfare from observed choices largely depends on the *values* that are judged to be important to conduct welfare analysis, which is an aspect that has largely been ignored in the literature. In the absence of a simple criterion that could identify the cases in which CS can be maintained, theorists need to be more explicit about the values they endorse to justify a certain characterisation of the individual's welfare. As sketched in the previous section when drawing a parallel with Arrow's (1951 [2012]) impossibility theorem, it does not seem possible to unambiguously integrate individual preferences across contexts into a single normative preference relation. This opens perspectives of further research about investigating (i) possible values regarding the definition of normative preferences, (ii) the compatibility between those values by studying the aggregation of conflicting intrapersonal preferences, and (iii) the investigation of intrapersonal bargaining. Many of those questions have extensively been studied in social choice theory, which suggests possible and promising bridges between behavioural welfare analysis and this literature.

A Appendix: Proofs

A.1 Proof of NCI

By contradiction, suppose that NCI is false and that there are two contexts γ and γ' such that $x \succ_\gamma y$ and $y \succ_{\gamma'} x$. By condition i. of NI, this means that there should be a context γ'' such that $x\ C_{\gamma''}\ y$ and another context γ''' such that $y\ C_{\gamma'''}\ x$, which violates CCI. This implies that NCI is true when both NI and CCI are true.

A.2 Proof of CS from CCI

By CCI we know that there are not two contexts γ and γ' such that $x\ C_\gamma\ y$ and $y\ C_{\gamma'}\ x$. This means that as soon as condition i. of NI is satisfied, so is condition ii. So if $x\ C_\gamma\ y$, we have $x \succ_\gamma y$. By CCI and NCI (which is implied by NI and CCI), we also know that the relation remains stable across all contexts γ and γ' for C and $\succ$, which means that $\succ_\gamma = C_{\gamma'}$.

References

Aivazian, V. A., J. L. Callen, and I. Lipnowski (1987). The Coase theorem and coalitional stability. *Economica 54*(216), 517–520.

Akerlof, G. A. and R. J. Shiller (2015). *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton University Press.

Arkes, H. R., G. Gigerenzer, and R. Hertwig (2016). How bad is incoherence? *Decision 3*(1), 20–39.

Arrow, K. J. (2012). *Social Choice and Individual Values* (third ed.). Yale University Press.

Atkinson, A. and J. Stiglitz (2015). *Lectures on Public Economics: Updated Edition*. Princeton University Press.

Bacharach, M. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton University Press.

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

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

Binder, C. (2014). Plural identities and preference formation. *Social Choice and Welfare 42*(4), 959–976.

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin (2003). Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism". *University of Pennsylvania Law Review 151*(3), 1211–1254.

Chambers, C. P. and T. Hayashi (2012). Choice and individual welfare. *Journal of Economic Theory 147*(5), 1818–1849.

Chetty, R. (2015). Behavioral economics and public policy: a pragmatic perspective. *American Economic Review 105*(5), 1–33.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics 3*, 1–44.

Dalton, P. S. and S. Ghosal (2011). Behavioral decisions and policy. *CESifo Economic Studies 57*(4), 560–580.

Dalton, P. S. and S. Ghosal (2012). Decisions with endogenous frames. *Social Choice and Welfare 38*(4), 585–600.

DellaVigna, S. (2009). Psychology and economics: evidence from the field. *Journal of Economic Literature 47*(2), 315–372.

Denett, D. (1987). *The Intentional Stance*. MIT Press.

Dietrich, F. and C. List (2013a). A reason-based theory of rational choice. *Noûs 47*(1), 104–134.

Dietrich, F. and C. List (2013b). Where do preferences come from? *International Journal of Game Theory 42*(3), 613–637.

Dietrich, F. and C. List (2016). Reason-based choice and context-dependence: an explanatory framework. *Economics and Philosophy 32*(2), 175–229.

Gao, X. S., G. W. Harrison, and R. Tchernis (2023). Behavioral welfare economics and risk preferences: a Bayesian approach. *Experimental Economics 26*(2), 273–303.

Gold, N. (2021). Guard against temptation: intrapersonal team reasoning and the role of intentions in exercising willpower. *Noûs 56*(3), 554–569.

Gonzalez, S., A. Marciano, and P. Solal (2019). The social cost problem, rights, and the (non)empty core. *Journal of Public Economic Theory 21*(2), 347–365.

Harrison, G. W., K. Morsink, and M. Schneider (2020). Do no harm? The welfare consequences of behavioural interventions. Technical report, CEAR Working Paper 2020.

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

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

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

Harrison, G. W. and D. Ross (2023). Behavioral welfare economics and the quantitative intentional stance. In G. W. Harrison and D. Ross (Eds.), *Models Of Risk Preferences: Descriptive And Normative Challenges*. Emerald.

Hédoin, C. (2015). From utilitarianism to paternalism: when behavioral economics meets moral philosophy. *Revue de Philosophie Économique 16*(2), 73–106.

Hershey, J. C. and P. J. H. Schoemaker (1985). Probability versus certainty equivalence methods in utility measurement: are they equivalent? *Management Science 31*(10), 1213–1231.

Hutt, W. H. (1940). The concept of consumers' sovereignty. *The Economic Journal 50*(197), 66–77.

Infante, G., G. Lecouteux, and R. Sugden (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.

Kahneman, D. (1999). Objective happiness. In D. Kahneman, E. Diener, and N. Schwarz (Eds.), *Well-being: The Foundations of Hedonic Psychology*, pp. 3–25. Russell Sage Foundation.

Kahneman, D., P. P. Wakker, and R. Sarin (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics 112*(2), 375–406.

Lecouteux, G. (2021). Behavioral welfare economics and consumer sovereignty. In *The Routledge Handbook of Philosophy of Economics*, pp. 56–66. Routledge.

Lecouteux, G. (2022a). Reconciling normative and behavioural economics: the problem that cannot be solved. In S. Badiei and A. Grivaux (Eds.), *The Positive and Normative in Economic Thought*, pp. 148–166. Routledge.

Lecouteux, G. (2022b). Reconciling normative and behavioural economics: The problem that cannot be solved. In *The Positive and Normative in Economic Thought*, pp. 148–166. Routledge.

Lecouteux, G. (2023). The homer economicus narrative: from cognitive psychology to individual public policies. *Journal of Economic Methodology 30*(2), 176–187.

Lerner, A. P. (1972). The economics and politics of consumer sovereignty. *The American Economic Review 62*(1/2), 258–266.

Loewenstein, G. and P. A. Ubel (2008). Hedonic adaptation and the role of decision and experience utility in public policy. *Journal of Public Economics 92*(8-9), 1795–1810.

Manzini, P. and M. Mariotti (2014). Welfare economics and bounded rationality: the case for model-based approaches. *Journal of Economic Methodology 21*(4), 343–360.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995). *Microeconomic Theory*. Oxford University Press.

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

Mitrouchev, I. (2019). Normative economics without the concept of preference. *Œconomia. History, Methodology, Philosophy 9*(1), 135–147.

Nozick, R. (1974). *Anarchy, State, and Utopia*. Blackwell.

Rizzo, M. J. and D. G. Whitman (2009). The knowledge problem of new paternalism. *BYU Law Review 2009*(4), 905–968.

Ross, D. (2005). *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.

Rubinstein, A. and Y. Salant (2012). Eliciting welfare preferences from behavioural data sets. *The Review of Economic Studies 79*(1), 375–387.

Salant, Y. and A. Rubinstein (2008). (A, f): choice with frames. *The Review of Economic Studies 75*(4), 1287–1296.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons.

Sen, A. (1970). The impossibility of a Paretian liberal. *Journal of political economy 78*(1), 152–157.

Sen, A. (2017). *Collective Choice and Social Welfare* (expanded ed.). Penguin Books.

Steedman, I. and U. Krause (1986). Goethe's Faust, Arrow's possibility theorem and the individual decision-taker. In J. Elster (Ed.), *The Multiple Self*, pp. 197–231. Cambridge University Press.

Sugden, R. (1985). Liberty, preference, and choice. *Economics & philosophy 1*(2), 213–229.

Sugden, R. (2004). The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. *American Economic Review 94*(4), 1014–1033.

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

Sugden, R. (2018b). 'Better off, as judged by themselves': a reply to Cass Sunstein. *International Review of Economics 65*(1), 9–13.

Sunstein, C. R. (2018). "Better off, as judged by themselves": a comment on evaluating nudges. *International Review of Economics 65*(1), 1–8.

Thaler, R. H. and C. R. Sunstein (2003). Libertarian paternalism. *American Economic Review 93*(2), 175–179.

Thaler, R. H. and C. R. Sunstein (2009). *Nudge: Improving Decisions about Health, Wealth, and Happiness* (revised and expanded ed.). Penguin Books.

Thoma, J. (2021). On the possibility of an anti-paternalist behavioural welfare economics. *Journal of Economic Methodology 28*(4), 350–363.

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*(4481), 453–458.

Tversky, A. and D. Kahneman (1986). Rational choice and the framing of decisions. *The Journal of Business 59*(4), S251–S278.

Tversky, A. and I. Simonson (1993). Context-dependent preferences. *Management Science 39*(10), 1179–1189.

Varian, H. R. (2014). *Intermediate Microeconomics: A Modern Approach* (ninth ed.). W.W. Norton & Company.