




ARTICLE

# Inferring welfare from inconsistent choices: how values matter

Guilhem Lecouteux<sup>1</sup> and Ivan Mitrouchev<sup>2</sup> 

<sup>1</sup>Université Côte d’Azur, CNRS, GREDEG, France and <sup>2</sup>Univ. Grenoble Alpes, INRAE, CNRS, Grenoble INP, GAEL, 38000 Grenoble, France

**Corresponding author:** Ivan Mitrouchev; Email: [ivan.mitrouchev@inrae.fr](mailto:ivan.mitrouchev@inrae.fr)

(Received 27 November 2023; revised 28 January 2025; accepted 06 February 2025)

## Abstract

There is no consensus on how to infer welfare from inconsistent choices. We argue that theorists must be explicit about the values they endorse to characterize individual welfare. After formalizing a set of values and their relationship with context-independent choices, we review the literature and discuss the advantages and drawbacks of each approach. We demonstrate that defining welfare a priori may violate normative individualism, arguably the most desirable value to maintain. To uphold this value while addressing individuals’ errors, we propose a weaker version of consumer sovereignty, which we label ‘consumer autonomy’.

**Keywords:** choice; context; preference; values; welfare

**JEL codes:** B41; D71; D90; I31

## 1. Introduction

Standard welfare economics is based on two fundamental premises. First, it is assumed that individuals make rational choices, in the sense that they satisfy a complete, reflexive, and transitive preference 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 to evaluate situations is the satisfaction of individuals’ preferences, as revealed by their choices (Varian 1987 [2014: Ch. 34]; Mas-Colell *et al.* 1995: Chs 16, 21). Assuming that people satisfy complete, transitive and continuous preferences, the theorist can represent individual choice as the maximization of a utility function.<sup>1</sup> Standard welfare economics then takes this

<sup>1</sup>By ‘theorist’ (she), we refer to the person – an economist, philosopher, expert or policymaker – who models the preferences of an ‘individual’ (he), and who may offer a normative judgement on the choice situation.

utility function as the individual welfare function (e.g. Graaff 1963).<sup>2</sup> However, evidence from behavioural economics challenges the first premise, which raises the question of how to define individual welfare based on preferences that do not necessarily satisfy rationality principles.<sup>3</sup> 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.<sup>4</sup> A related approach consists in estimating the most likely preference relation that can rationalize choice data, and using deviations from this preference relation as a measure of welfare loss (Apesteguia and Ballester 2015; Echenique *et al.* 2023). Although the literature is substantial and still growing, there is currently no consensus on how to infer welfare from possibly inconsistent choices.

We propose to make explicit the *values* that theorists endorse when providing welfare evaluation from possibly inconsistent choices, and to discuss the desirability of these values. By ‘values’, we refer to the principles held by the theorist when forming judgements about the *normative* preferences of individuals.<sup>5</sup> We propose a set of values that characterizes the relationship between individual choice and welfare: (i) normative individualism, (ii) normative context-independence and (iii) consumer sovereignty. We then discuss the relationship between these values and the condition of (iv) choice context-dependence, which is implicit in standard welfare economics, and which allows for an unambiguous definition of the individual welfare function. We show that since behavioural economics challenges the validity of (iv), theorists must explicitly rely on (i), (ii) or (iii) in order to derive the individual welfare function. Based on our formal characterization of (i), (ii), (iii) and (iv), we then review the main approaches in the literature: the choice-based framework (Bernheim and Rangel 2007, 2009), behavioural paternalism (Thaler and Sunstein 2003, 2009), quantitative intentional stance (Harrison and Ross 2018), opportunity (Sugden 2004, 2018a), and experienced utility (Kahneman *et al.* 1997). A key distinction between these approaches is that they rely on different values (i.e. conditions to define individual welfare) and different theories of ‘errors’ (i.e. conditions to define individual deviations from welfare maximization). After arguing that investigating errors is crucial to develop normative evaluations in the presence of inconsistent choices, we propose an additional value of *consumer autonomy*, which derives from a characterization of errors we develop in Lecouteux and Mitrouchev (2024).

<sup>2</sup>In choice under risk, utility is traditionally used to designate the Von Neumann–Morgenstern utility of outcomes, and the utility of a prospect is characterized as the subjective expected Von Neumann–Morgenstern utility of the outcomes of the prospect. In this paper we focus on preferences, where utility is defined over alternatives, not over outcomes.

<sup>3</sup>See McQuillin and Sugden (2012) and Chetty (2015) for overviews from different perspectives of this challenge and Mitrouchev (2024) for a state of the art. For a literature review of empirical deviations from the standard model of rational choice, see DellaVigna (2009).

<sup>4</sup>See Bernheim and Rangel (2007, 2009), Salant and Rubinstein (2008), Dalton and Ghosal (2011, 2012), Chambers and Hayashi (2012), Rubinstein and Salant (2012), Manzini and Mariotti (2014). For a discussion, see Bernheim (2016) and Thoma (2021).

<sup>5</sup>This question has been extensively discussed in standard welfare economics, with, e.g. Mongin’s (2006) investigation into value judgements and value neutrality. One of our objectives is to offer an explicit representation of the question.

The rest of the article is organized as follows. We first define a formal framework that characterizes how ‘normative’ preferences can be derived from inconsistent choices, along with the values (i), (ii), (iii), and the condition of choice context-independence (iv) (section 2). Within this framework, we review the main approaches in the literature (section 3). We then discuss the limitations of each approach, highlighting the respective values they endorse and/or reject, as well as the challenge of maintaining normative individualism if welfare is defined *a priori*. In response to this challenge, we propose a value of ‘consumer autonomy’ as a weaker form of consumer sovereignty (section 4). Our main conclusion is that, in the absence of a simple criterion to identify cases where consumer sovereignty can be maintained, theorists must be more explicit about the values they endorse to justify a particular characterization of individual welfare. This requires placing greater emphasis on the characterization of *errors* rather than on welfare (section 5).

## 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 (Bernheim and Rangel 2007, 2009; among others). 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.<sup>6</sup> 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 (e.g. colour, shape) or extrinsic properties of the alternatives (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 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

<sup>6</sup>Our definition of context is therefore quite 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).

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 characterizing Smith, which could be represented as follows.

- $P_b(\text{Smith}) = 1$ . This means that the property ‘boldness’ is satisfied.
- $P_p(\text{Smith}) = 0.5$ . This means his political agenda, on a range of real numbers from 0 (far left) to 1 (far right) is in the centre.
- $P_v(\text{Smith}) = 80$ . This represents the score of his visibility on social media, from 0 to 100.
- $P_m(\text{Smith}) = 1$ . This means 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 boldness and political agenda, but considers that only the political agenda is relevant for his vote. Moreover, he does not know that Smith is a manipulator, which should – at least from the perspective of the theorist – also be relevant for his vote (as Smith may not be trustworthy). Furthermore, he does not know that social media visibility – which is not relevant to his vote – may influence his decision. Here, we have a situation in which one property is relevant, motivational and known (Smith’s political agenda), another which is relevant but neither motivational nor known (Smith’s manipulation), another which is motivational but neither known nor relevant (Smith’s visibility), and finally, a property which is known but neither relevant nor motivational (Smith’s boldness).<sup>7</sup>

Our definition of the context is based on the premise that it refers to what *we* theorists consider 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 relevant.<sup>8</sup> For simplicity, we assume that the theorist correctly identifies the set  $\mathcal{M}_I$ , i.e. she knows precisely the properties that influence the choice of the individual.<sup>9</sup> Formally, a *context property* is a property that is motivational but not relevant:  $P \in \mathcal{C}_I = \mathcal{M}_I \setminus \mathcal{R}_I$ . A context is any combination  $\gamma = (\gamma_P)_{P \in \mathcal{C}_I} \in \Gamma$  of values of the context properties. In the example above, there is only one

<sup>7</sup>We could have expanded this illustration with other cases, e.g. motivational and known, but irrelevant 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  $\mathcal{M}_I$ ,  $\mathcal{K}_I$  and  $\mathcal{R}_I$ .

<sup>8</sup>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 Lecouteux and Mitrouchev (2024).

<sup>9</sup>Relaxing this assumption would lead us to consider that the theorist could have an incorrect representation of the choice problem, a complication we prefer to avoid.

property – visibility on social media – that is motivational and not relevant, i.e.  $C_I = \{P_v\}$ , 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  $X$ .<sup>10</sup> 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  $\gamma$ , and denote it  $C_\gamma \subset X \times X$ . We interpret  $C_\gamma$  as a choice ranking: ' $x C_\gamma y$ ' reads as ' $I$  chooses  $x$  over  $y$  in context  $\gamma$ '. It means that, when asked to choose between  $x$  and  $y$  in a context  $\gamma$ ,  $I$  chooses  $x$ . We do not make any assumption about the properties of  $C_\gamma$ , 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  $\gamma$ . It is the ranking that characterizes the individual's welfare.<sup>11</sup> While  $C_\gamma$  represents the actual choice of the individual in context  $\gamma$ ,  $\succ_\gamma$  represents the preference that he ought to satisfy in order to maximize his welfare. The distinction between  $C_\gamma$  and  $\succ_\gamma$  allows us to differentiate between the 'descriptive' and 'normative' aspects of individual decision-making. For convenience, assume that  $C_\gamma$  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 could be to define the normative preferences of an individual as the preferences he would reveal if he were 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).

<sup>10</sup>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.

<sup>11</sup>Various terms are used in the literature, including true, authentic, laundered and implicit preferences (among others). Our concept of normative preference is intentionally broad, meaning we do not specify a particular *type* of preference that makes the individual better off, such as one conforming to specific principles of rational choice. Furthermore, while we, as authors, remain fundamentally agnostic about the precise definition of welfare, our framework adopts the preference satisfaction view. We believe this perspective aligns well with respecting individual autonomy, as it recognizes that individuals are the best judges of what benefits them, thereby avoiding the imposition of external standards of what constitutes the 'good'. This view underpins our defence of 'Normative Individualism' and the 'Consumer Autonomy' value we propose in section 4.

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 asked to choose between two different health programmes, represented by a certain and a risky alternative. The choice between the two programmes can be framed in terms of gains or losses.<sup>12</sup>

*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%]<sup>13</sup>

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 difficult to identify individuals' normative preferences.

### 2.3 Values and Context-Independence

One way to clarify the question of inferring welfare from inconsistent choices is to distinguish between the *theory of welfare* and the *theory of error* endorsed (implicitly or not) by the theorist. A theory of welfare corresponds to the framework used by the theorist to define individual welfare, i.e. the set of *values* she endorses. The theory of error corresponds to the framework used by the theorist to define the individual's deviation from welfare maximization, i.e. the deviation caused by the *context*. Our first goal is to review the literature in 'behavioural' normative economics by highlighting the implicit values that are endorsed in different contributions, and to point out the incompleteness of certain approaches, which may lack a clear theory of error. We characterize three main values: (i) normative individualism, (ii) normative context-independence and (iii) consumer sovereignty.

<sup>12</sup>The % below corresponds to the share of subjects who chose the programme in the experiment, and N corresponds to the total number of subjects per frame.

<sup>13</sup>This experiment is a survey response based on an unincentivized hypothetical choice task. Yet we could have referred to other examples with an incentivized choice task, such as known discrepancies between different preference elicitation methods, e.g. certainty equivalence and probability equivalence (Hershey and Schoemaker 1985). The reason for choosing this example is that it offers a simple and clear illustration of how the language chosen by the experimentalist can change the perception of individuals, and ultimately their stated preferences. It also originally refers to the concept of *framing*, as coined by Tversky and Kahneman (1981).

According to normative individualism, the proper locus of normative concern is individual persons, whose own values and situations should be taken into account when debating ethical issues such as policy or justice.<sup>14</sup> We translate this value in our framework as follows.

**VALUE 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  $\gamma'$  such that  $x C_{\gamma'} y$
- ii.  $x \succ_\gamma y$  if  $x C_{\gamma'} y$ ,  $\forall \gamma' \in \Gamma$

This value establishes a close relation between the choice and the normative preference of the individual.  $x$  can be considered better than  $y$  in context  $\gamma$  only if there exists at least one context  $\gamma'$  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  $x$  is necessarily better than  $y$  (condition ii). The main idea 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, NI remains silent on cases where the choice between  $x$  and  $y$  depends on the context. This is, however, not true of the following values, namely (ii) normative context-independence and (iii) consumer sovereignty.

**VALUE 2. 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 value 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 were not the case, the welfare associated with a given alternative could vary depending on the context of choice, resulting in a welfare function that is unstable across contexts. An alternative value on which we can rely to infer the individual's welfare is:

**VALUE 3. 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.<sup>15</sup> More specifically, this value states that the

<sup>14</sup>See Ross (2005: 220–222) for a contemporary definition. This value has obviously deeper ideological and philosophical roots that could be found in foundational references such as J.S. Mill (1849, 1859).

<sup>15</sup>This concept was originally formulated by Hutt (1936), then reformulated by himself 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 predominantly 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).



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.

While NI is usually too general to allow a single characterization of the individual’s welfare function, this is not the case for NCI and CS, which, however, do not lead to the same characterization. CS allows for context-dependent normative preferences, which is excluded by NCI. We can indeed note several conditions of inclusion and compatibility between these values. First, CS is more restrictive than NI. CS respects condition (i) of NI by construction, although it imposes that  $\succ_\gamma$  necessarily corresponds to  $C_\gamma$  (while according to NI,  $\succ_\gamma$  is known for sure only if the choice between two alternatives remains 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), then CS implies  $\succ_\gamma \neq \succ_{\gamma'}$ , which leads to a violation of 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  $x C_{\gamma'} y$ . We suggest that this indeterminacy is addressed in standard welfare economics thanks to an implicit condition of choice context-independence, which implies an absence of ‘error’.

**CONDITION 1. Choice Context-Independence (CCI).**  $\forall \gamma, \gamma' \in \Gamma, C_\gamma = C_{\gamma'}$ <sup>16</sup>

CCI states that  $I$ ’s choice does not depend on the context in which he is embedded. According to our framework, this means that the set of context properties is empty. Any motivational property is necessarily relevant, and vice versa. It can easily be shown that, if CCI is verified, NI implies both NCI and CS (Appendix A). This means that we can always and unambiguously define the normative preferences of the individual – which necessarily equate to his preferences  $C_\gamma$ . 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, CCI is a tautology (choices do not depend on the context). CS therefore holds in standard welfare economics, as well as NCI. The challenge raised by behavioural economics is, however, that CCI does not hold in many situations. This means that NCI or CS must be *postulated* in order to derive normative preferences. Furthermore, the characterization of normative preferences that is derived from NCI is not compatible with the characterization derived from CS anymore. That is, the welfare function that would be inferred by maintaining NCI is different from the one that would be inferred by maintaining CS. This means that the theorist must choose one of these two values before eliciting normative preferences.

### 3. Literature Review

We now discuss in detail the main alternatives to derive normative preferences when CCI does not hold. We categorize the literature as follows: choice-based framework, behavioural paternalism, quantitative intentional stance, opportunity and experienced utility.

<sup>16</sup>Since choice context-independence does not refer to the normative preferences of the individual, it is not a value. We formulate it as a condition (to be verified or not).



### 3.1. Choice-Based Framework

The choice-based framework (Bernheim and Rangel 2007, 2009) consists in extending 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, which 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 treated 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 making normative evaluations 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 acknowledged 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 a pragmatic strategy to the challenge of inferring welfare from inconsistent 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 the 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 situations where individual choice is ‘too’ inconsistent.

Thoma (2021) notes that the choice-based framework might be silent in some situations because ‘the agent’s underlying desires and their respective importance may simply not be precise enough to determine one unique and complete integrated preference relation that correctly aggregates them’ (360). This limitation is well recognized by Bernheim (2016), who argues that in such cases, ‘it is important to acknowledge that our inability to make precise normative statements reflects the limits of our knowledge’ and that ‘admitting this ambiguity is intellectually honest’ (60).<sup>17</sup> This means that the range of situations that can be studied is rather restricted, and the theorist cannot conduct welfare analysis in situations where choices vary highly across contexts (while those are potentially highly relevant for policy).<sup>18</sup>

The limitation of the choice-based framework seems to be its inability to offer an unambiguous theory of error, i.e. a clearly identified framework that would help the theorist to identify a priori which choices are erroneous. This difficulty derives from the lack of a unified paradigm in cognitive psychology and the tendency of

<sup>17</sup>This echoes a point that we voluntarily left aside in this paper (see footnote 9, i.e. what to do when the theorist is in an impoverished epistemic position). Questioning the epistemic position of the theorist turns out to be crucial when looking for an adequate approach to design public policies (Lecouteux 2021b: 224–226).

<sup>18</sup>Lecouteux (2021b) argues that the set of situations that can be studied under this framework is akin to *microcosms* in the sense of Savage (1954).

behavioural economists to document the accumulation of ‘biases’ without providing a common framework from which we could systematically derive how context properties influence individual choices. Bernheim (2016) acknowledges the difficulty of defining what a ‘mistake’ is (apart from obvious cases, such as crossing the street in the UK while looking at the wrong side of the road). He proposes two defining features for a ‘mistake’. It is inconsistent with the information available to the decision maker, and the individual would have chosen another option if the ‘characterization failure’ had not occurred. Given Bernheim’s characterization, it is hard to ascertain, for instance, whether any behaviour that generates a risk for one’s health (e.g. eating too much sugar or salt, or drinking alcohol) could be considered a mistake. Given the evidence in medical sciences, everyone *should* stop drinking alcohol if they want to preserve their health. However, this characterization might seem at odds with the relatively liberal inspiration of the choice-based framework, according to which the theorist should impose no constraints on people’s preferences.<sup>19</sup>

### 3.2. Behavioural Paternalism

Behavioural paternalism characterizes individual welfare as the satisfaction of preferences when people’s decisions are not distorted by cognitive biases.<sup>20</sup> 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 into our framework, CCI is here explicitly rejected, while NI is intended to be maintained.<sup>21</sup> Here, the rejection of CCI leads to the rejection of CS, since it is considered that individuals can make mistakes, while NCI is maintained, i.e. the adequate context to infer normative preferences is when the individual is not influenced 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 *et al.* (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.

<sup>19</sup>See in particular Bernheim (2016: 17–18), who distinguishes between direct and indirect judgements when an individual must choose between different alternatives. According to Bernheim, direct judgements are what the individual thinks is good for himself, while indirect judgements concern what the individual thinks he should do to achieve what is good for himself. In Bernheim’s words, ‘there is nothing wrong with direct judgements’ (18), and the theorist is in no epistemic position to make ethical judgements about those.

<sup>20</sup>This is the theoretical approach to welfare in behavioural paternalism, which is the one we discuss here. In practice, behavioural paternalism rather *exploits* people’s biases to guide them towards the desired behaviour (e.g. eating more healthily). 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 individuals with no (or minor) cost to rational individuals.

<sup>21</sup>In this literature, NI corresponds 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 of satisfying this clause.

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 an acceptable choice rule for the individual. In this case, normative preferences would be considered context-dependent as well, which eventually leads to a violation of NCI. As a result, it may not be possible to define a stable (context-independent) welfare relation from individuals' 'de-biased' choices.

Second, it is not obvious that the theorist can correctly identify the context properties that are motivational but not relevant.<sup>22</sup> 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. Consider 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 be perfectly fine with it – e.g. he considers it 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 (Tversky and Kahneman 1981), the theorist cannot know a priori whether the individual ought to be risk-averse or risk-seeking. This suggests that NI is not necessarily satisfied in behavioural paternalism, despite the narrative promoted by proponents of this literature. Indeed, behavioural paternalism imposes *consistency* across contexts as a normative criterion, which appears to be more controversial than is usually considered and would require additional justification.<sup>23</sup>

Behavioural paternalism shares a common limitation with the choice-based framework: the absence of an unambiguous theory of error. Its strategy to infer welfare from inconsistent choices involves imagining the counterfactual preferences of an inner rational agent, free from contextual influences. However, contrary to its proponents' claims, this requires setting arbitrary conditions for defining individual welfare, such as a condition of consistency across contexts, like NCI. A further complication is that NCI is insufficient to offer a clear, unambiguous definition of individual welfare. This means theorists must introduce additional conditions on normative preferences. For example, in much of the intertemporal choice literature, it is often assumed that individuals would prefer to be more patient and place greater weight on future consequences, though this is likely to reflect the *theorist's own preferences* rather than those of the individuals she models.

### 3.3. Quantitative Intentional Stance

This approach proposed by Harrison and Ross (2018, 2023) is based on Dennett's (1987) externalist account of preferences and beliefs. These are not defined as inner

<sup>22</sup>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 unknown in public economics, where a fundamental task of the theorist is to set up an incentivized mechanism so that individuals reveal their 'true' preferences (Atkinson and Stiglitz 2015: Ch 16.6). In this framework, however, the problem is rather one of *trustworthiness* between the theorist and individuals than of welfare elicitation per se.

<sup>23</sup>See Arkes *et al.* (2016) and Lecouteux (2021a) for an extensive analysis of the lack of normative justification for consistency.

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 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) characterize 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.<sup>24</sup> 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 *et al.* 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 retains NI, as well as NCI. The suggestion that 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 are minimized. The relative arbitrariness of the definition of welfare, as the most likely (econometrically speaking) utility structure characterizing the individual preferences and beliefs, is explicitly recognized here as the theorist's prior. There is, therefore, a possibility of 'mistake' (Harrison and Ross 2023: Ch. 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 bank employee whose aim is to improve his clients' financial choices). 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 with the choice-based framework – offers an operational approach to welfare analysis by being explicit about its theory of error,

---

<sup>24</sup>This is because such experiments are considered '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 characterization (e.g. taking the mean to estimate the parameters) of the risk preferences and beliefs (Gao *et al.* 2023). Unlike 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.

with a clear framework to elicit individual welfare from choice data. However, it 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 by rejecting NCI and shifting the normative focus from welfare to *opportunity*. This strategy values individual *freedom* of choice rather than actual choices. The role of the theorist is not to make policy recommendations that maximize individual welfare, but to ensure that institutions are designed in such 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 maximizes the opportunity sets of its participants and thus facilitates the pursuit of mutual benefits – in which case the market is seen as a cooperative rather than a competitive institution (Sugden 2018a). Unlike the rest of the literature discussed in this article, the theorist has no role in identifying the relevant properties of a choice situation, as she does not aim to make normative evaluations from individuals’ preferences at all.<sup>25</sup> 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 into 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 provides 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 self-acknowledged failures of self-control (e.g. drug addiction) and perhaps most importantly, situations where individuals’ preferences are strongly influenced by unknown properties (e.g. aggressive marketing or adaptive preferences) – whose knowledge may result in changing their choices. One example of a 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) from 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 outright forms of fraud and deception on the part of firms. This could, however, result in violating the rules of fair competition, that each individual is initially expected to accept. Maintaining CS in all circumstances means there is no room for a theory of error, which seems too strong.

<sup>25</sup>See Mitrouchev (2019) for a detailed assessment of the opportunity approach compared with behavioural paternalism.

### 3.5. Experienced Utility

Yet another alternative to infer welfare from inconsistent choices is to consider people's level of *hedonic states* (pain and pleasure) for welfare evaluation. This is captured by the concept of *experienced utility* (Kahneman *et al.* 1997). In contrast with decision utility, which refers to the weight given to an alternative in a decision (and which is therefore based on individuals' choices), experienced utility refers to the actual experience of choosing one alternative over another, in line with the Benthamite pain/pleasure dichotomy. In this approach, it is explicitly acknowledged that decision utility is context-dependent, meaning that CCI is rejected. In our framework, hedonic states (pain and pleasure) are derived from the satisfaction of normative preferences, which satisfy the conditions (i or ii) of NI. Thus, experienced utility intends to hold NI (we, however, challenge this possibility below). Contrary to behavioural paternalism and the choice-based framework, this approach avoids relying on extra criteria (such as consistency) since the preference relation is cardinally defined for similar types of choices.<sup>26</sup> For example, an individual can derive more pleasure from consuming ice cream than a doughnut in summer, while he can derive more pleasure from a doughnut over ice cream in winter – these two hedonic states being directly comparable because they refer to commensurable 'units' of happiness. Allowing for the possibility that individuals' normative preferences are context-dependent, NCI is then rejected.

There is yet a particular rule for how people's hedonic states are aggregated. In particular, Kahneman (1999) attempts to specify what an external observer would need to know to determine how happy an individual is at a given time, along with the rules for using that knowledge. According to Kahneman, the highest level of evaluating welfare is grounded in information about *moment utility*: what is experienced here and now by the individual. Kahneman distinguishes two notions of happiness: *subjective happiness*, based on self-reported measures ('how happy are you?' Likert scales) and *objective happiness*, which is derived from a record of moment utility over the relevant period. He emphasizes that remembered utilities (what is remembered of an experience) and total utility (the aggregation of all hedonic states experienced at each moment) of episodes differ, much like subjective and objective happiness. In his view, the former gives an approximate evaluation of one's welfare (here, happiness), while the latter provides a more precise measure of happiness.<sup>27</sup> Although objective happiness is based on subjective self-reports, the aggregation of moment utility is governed by a rule *external* to the individual, i.e. one imposed by the *theorist*.<sup>28</sup> From Kahneman's (1999) viewpoint, only *objective* happiness is normatively relevant. This has strong implications for CS. As the author argues, 'policies that improve the frequencies of good experiences and reduce the incidences of bad ones should be pursued *even if people do not describe*

<sup>26</sup>See in particular the *cardinality of instant utility* axiom in Kahneman *et al.* (1997).

<sup>27</sup>This normative stance arises from Kahneman's experiments with colleagues, where they found that people not only fail to optimise moment utility (Redelmeier and Kahneman 1996), but also fail to choose the experience that minimizes pain when asked to choose between two repeated experiences (Kahneman *et al.* 1993).

<sup>28</sup>This is part of the normative theory of Kahneman *et al.* (1997) and explicitly acknowledged in Kahneman (1999).

*themselves as happier or more satisfied* (15 – our emphasis). In this regard, CS is strongly rejected.

One significant limitation of this approach is that rejecting CS leads to a full delegation of individual welfare to the theorist.<sup>29</sup> Following Kahneman's line of reasoning, experienced utility leads to even more paternalistic interventions than behavioural paternalism, imposing great restrictions on the normative value of individuals' self-reports. This also raises questions about NI. In particular, there is no clear justification for why the aggregation of *moment utility* should be what ultimately matters for the individual's total utility. In fact, Kahneman himself revised his stance on objective happiness later in his career, arguing that life satisfaction in terms of what people *remember* of their past experience may matter more than the maximization of their moment utilities.<sup>30</sup> There is indeed no particular reason why moment utility should be given priority over remembered utility, even if remembered utility may contradict the principle of moment utility maximization. After all, one could argue that what truly matters is *the individual's perception of his own experiences*, rather than an externally imposed aggregation rule (i.e. one dictated by the theorist). That is, one may prefer to assign normative value to remembered utility if one considers that the individual's memory is what matters *to him*. By arbitrarily privileging moment utility (rather than, for instance, remembered utility), experienced utility also lacks a clear and unambiguous theory of error.

## 4. Errors and Autonomy

### 4.1. Comparison of The Different Approaches

Table 1 summarizes the positions of the approaches we reviewed in section 3. A checkmark means that the value or condition is maintained. A crossmark means that the value or condition is rejected.

We put a crossmark for NI and behavioural paternalism, as the *intention* of behavioural paternalism is to uphold NI. However, this condition is unlikely to be met in practice due to the absence of a clear theory of error and the imposition of a criterion of consistency on one's true preferences. A similar concern applies to experienced utility, where NI is theoretically satisfied if one accepts that moment utility should form the informational basis of the theory. However, as previously discussed, this assumption is far from self-evident. Because experienced utility also lacks an unambiguous theory of error, it is difficult to assert that experienced utility genuinely respects NI. We therefore put a crossmark.

<sup>29</sup>It is important to note, however, that compared with other approaches, experienced utility is the only one that grounds normative evaluation in a well-defined ethical theory – namely, *Benthamian hedonism*. This has both merits and drawbacks. It explicitly acknowledges that one cannot compare individual (consequently social) situations without taking a clear stance on what constitutes the good – a position to which we are sympathetic and that aligns with our proposition to be more explicit about one's values. A drawback, however, is that not everyone may agree with such an ethical theory, particularly due to its reductionist nature.

<sup>30</sup>See the full interview of Daniel Kahneman by Amir Mandel in March 2018 for *Haaretz* newspaper: <https://www.haaretz.com/israel-news/premium/MAGAZINE-why-nobel-prize-winner-daniel-kahneman-gave-up-on-happiness-1.6528513h>.



**Table 1.** Value/condition-check of literature review

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

We can also observe that all approaches, except experienced utility, uphold either NCI or CS. This aligns with the necessity of postulating NCI or CS to derive normative preferences. Our analysis suggests that rejecting CCI leads to an incompatibility between NCI and CS, which can only be reconciled by remaining silent on ambiguous choice data, as seen in the choice-based framework. Furthermore, if the theorist aims to provide welfare measures from his own perspective, NCI must be maintained. The individual's welfare can then either be inferred from choice data by explicitly incorporating a theory of error (as seen in the quantitative intentional stance, where errors are treated as noise), or by imposing the normative preferences of the inner rational agent, as in behavioural paternalism. The opportunity approach rejects the possibility of error altogether and shifts the normative focus from welfare to opportunity. As for experienced utility, it presents a potential solution to the challenge of measuring welfare without adhering to either NCI or CS. However, due to its lack of a clear theory of error, it leaves NI unspecified.

Two alternative paths seem possible for inferring welfare from inconsistent choices. The first is that we, as theorists, start from choice data (respecting NI) and are explicit about the theory of error – whether it involves restricting choice data in the choice-based framework, treating errors as noise in the quantitative intentional stance, or dismissing errors entirely in the opportunity approach. The second is that we begin with an a priori definition of welfare (such as the satisfaction of true preferences in behavioural paternalism or the maximization of hedonic states in experienced utility) and evaluate situations based on this ethical judgement. A caveat with this second alternative is that it risks a paradox regarding NI, as seen in experienced utility, which leaves room for interpretation about what truly matters to individuals. Moreover, some may argue that defining welfare based on an ethical theory undermines liberal principles, where individuals should be free to support their own ethical views.<sup>31</sup>

In fact, an approach that relies on a definition of welfare is likely to violate NI. As discussed earlier, NI remains silent in cases where choice is context-dependent. This suggests the need to identify general values that define desirable properties of normative preferences and their relationship to choices across different contexts.

<sup>31</sup>This reflects a problem of value incompatibility, similar to those in social choice theory. Sugden (2018a) highlights this parallel between behavioural welfare analysis and social choice theory in the preface of *The Community of Advantage* (viii–ix), where he references his critique of Sen's impossibility of a Paretian liberal (Sen 1970; Sugden 1985) and his proposition of the individual opportunity criterion (Sugden 2004).

The second condition of NI illustrates this:  $x \succ_{\gamma} y$  if  $x C_{\gamma'} y$ ,  $\forall \gamma' \in \Gamma$ . This resembles the unanimity condition in social choice theory, where if all individuals prefer  $x$  to  $y$ , then  $x$  must be socially preferred. In fact, many paradoxes or impossibility theorems in social choice theory, such as those of Arrow (1951 [2012]) and Sen (1970, 2017), can be transposed to the intrapersonal level if we treat an individual as a collection of subpersonal selves defined over different contexts. For example, desirable properties for normative preferences could mimic those of Arrow's (1951 [2012]) impossibility theorem.<sup>32</sup>

- **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.**<sup>33</sup> 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.** There is no  $\gamma^* \in \Gamma$  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: we can note that this condition extends the sets of preferences that must be taken into account in welfare analysis (with the introduction of context-dependence), which can be interpreted as the violation of CCI, and which gives rise to the challenge of inferring welfare from inconsistent choices. This condition is verified with experienced utility and behavioural paternalism, but not with opportunity (since  $C_{\gamma}$  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 except experienced utility. Non-dictatorship is verified in the opportunity approach, while being rejected in behavioural paternalism, which imposes choice in a 'context-free' situation as the legitimate one.<sup>34</sup>

From this overview of the different approaches with respect to the values we consider, we can see that almost all approaches that maintain NCI reject at least one of the other values.<sup>35</sup> From a methodological point of view, the problem of preference *integration* is closely related to the problem of preference *aggregation* in

<sup>32</sup>In what follows, we drop the subscript  $\gamma$  for  $\succ_{\gamma}$ , since we must respect NCI in order to define a welfare function.

<sup>33</sup> $\langle C_{\gamma'} \rangle(x; y)$  denotes the ranking between  $x$  and  $y$  induced by the choice functions  $\{C_{\gamma'}\}_{\gamma' \in \Gamma}$ .

<sup>34</sup>The relationship between independence of irrelevant alternatives and the various approaches reviewed earlier is less straightforward, which is the reason we do not discuss it further.

<sup>35</sup>The exception is the choice-based framework, which aims to satisfy all values and conditions. However, as previously discussed, this approach results in a restricted scope for normative analysis.

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. In the literature addressing this point, Steedman and Krause (1986) and Binder (2014) characterize the conditions under which aggregation is possible at the intrapersonal level. In a nutshell, they suggest that aggregation may only be possible if the degree of conflict between the various choices of the individual is low. Defining the welfare of the individual from a priori principles might thus lead to a violation of NI, and it seems that integrating individual context-dependent preferences into a single stable normative preference might be particularly challenging. For this reason, a more pragmatic approach, in our view, involves investigating the ‘right’ *theory of error* rather than the ‘right’ theory of welfare.

#### 4.2. Consumer Autonomy

We propose that NI should constitute the basis of behavioural welfare analysis. That is, welfare evaluation should ultimately depend on the individual’s *choices*, even though we should recognize the possibility of errors (i.e. there may exist  $\gamma \in \Gamma$  for which  $C_\gamma \neq \succ_\gamma$ ). As discussed in the previous section, maintaining NI seems, however, to be incompatible with an integrative approach to welfare (as it would likely lead to a violation of one of the conditions listed above). The alternative is then to consider the conditions under which the individual makes *errors*, i.e. the mechanisms through which his choices are influenced by the context.

In section 3, we argued that rejecting the possibility of error (i.e. endorsing CS as a stronger formulation of NI, as in the opportunity approach) might be problematic in certain situations – e.g. self-acknowledged failures of self-control or addictions – and that we need to consider that some motivational properties might not be relevant. The challenge we face is, however, that psychology lacks a unified framework to characterize erroneous choices. Our proposition – which we develop more extensively in Lecouteux and Mitrouchev (2024) – is 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. We label this approach the ‘view from *Manywhere*’. This confrontation of perspectives respects NI while not taking CS *prima facie*. Normative preferences are indeed fundamentally related to individual choices, while we recognize the possibility of errors – i.e. choices made in certain contexts that, viewed from the perspective of another context, are not accepted by the individual. Indeed, since errors are akin to contextual properties in our framework, confronting views from different contexts can help the individual become aware of those contextual properties, and possibly prevent their influence on the final choice.<sup>36</sup>

<sup>36</sup>For example, in Tversky and Kahneman’s (1981) Asian disease experiment, being able to perceive the problem *both* in terms of gains and in terms of losses is likely to protect the individual from a pure framing effect, as observed in the actual experiment (a tendency to be risk-averse in the gain frame and risk-seeking in the loss frame). In our framework, the ‘error’ is corrected by the fact that the individual becomes aware of the influence of context properties.

Investigating errors could then provide normative guidance to *avoid* such errors. By improving the process through which individuals form their preferences, we could define conditions under which individuals may choose to ignore (or not) context properties. This would shift the normative focus from welfare and the satisfaction of individual preferences to *autonomy* and the process of preference formation.<sup>37</sup> Formally speaking, it would be possible to investigate these questions by referring to debates in social contract theory (rather than preference aggregation) applied to intrapersonal bargaining.<sup>38</sup> This would lead to the formulation of procedural normative criteria, under which we could confidently maintain CS. The non-respect of such criteria, however, in the presence of e.g. manipulations by a third party, would help the theorist to identify potential context properties, and thus, errors. In order to accommodate CS, which, in our view, is too strong, we formulate a weaker version of CS as follows:

**VALUE 4. Consumer Autonomy (CA).**  $\forall \gamma \in \Gamma, C_\gamma = \succ_\gamma$  only if  $\mathcal{C}_\gamma \subseteq \mathcal{K}_\gamma$

The value means that, if some context properties (influential yet non-relevant properties) are unknown to the individual, then we cannot systematically follow CS. It is indeed possible that the choices of the individual are caused by factors that he would consider as irrelevant upon careful scrutiny. Furthermore, if CCI holds, then this value is identical to CS by construction as  $\mathcal{C}_\gamma = \emptyset$ . Note, however, that being fully aware of all context properties does not automatically imply CS. A situation of self-acknowledged failure of self-control would not qualify as an autonomous choice, since the individual would like not to be influenced by the context, but cannot do otherwise. Being aware of the context properties constitutes then a necessary condition to make autonomous choices. A complementary condition – which cannot directly be transposed into our framework, in the absence of a model on intra-personal bargaining – is one of *authenticity*, i.e. that the individual accepts that his choice is influenced by some context properties (see Christman 2009; Lecouteux 2022). In terms of policy guidance, this means that the aim of the theorist is to make the individual aware of more context properties, rather than trying to infer counterfactual normative preferences when the individual might be influenced by context properties of which he is not aware (see Lecouteux and Mitrouchev (2024) for a more in-depth discussion). Our approach therefore rejects CCI, NCI and CS, while respecting NI, as well as CA (#Table 2).

<sup>37</sup>See in particular Lecouteux (2022) on definitions of autonomy in ‘behavioural’ normative economics.

<sup>38</sup>For instance, Hédoin (2015) argues that ‘behavioural economists have totally ignored the solution of Coase (1960), which consists in letting the individual’s various selves (interpersonally) bargain over the internalities’ (78). If assumptions about bargaining between individuals make sense when transposed to bargaining between selves, some results on *social* bargaining could likely be transposed to *individual* bargaining. For example, 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 met more easily (Aivazian *et al.* 1987). We can also imagine that the problem could be addressed with the tools of cooperative game theory (Gonzalez *et al.* 2019), or with a model of intrapersonal team reasoning (Gold 2022).

**Table 2.** Value/condition-check of our proposition

	CCI	NI	NCI	CS	CA
View from <i>Manywhere</i>	X	✓	X	X	✓

## 5. Conclusion

There is no consensus on how to infer welfare from inconsistent choices. We argue that the different approaches proposed in the literature rely on various values endorsed by theorists regarding the relationship between inconsistent 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 is, first and foremost, the theorist’s representation. We identified three values that characterize the structure of normative preferences: (i) normative individualism, (ii) normative context-independence and (iii) consumer sovereignty. Standard welfare economics does not consider the possibility of context properties (i.e. properties of the alternatives that are motivational but not relevant). In our framework, this means that (iv) choice context-independence is assumed. The direct consequence is that both NI and CS have the same characterization 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 strictly maintain CS (as in the 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 relevant, although we may encounter disturbing cases (e.g. addictions and deceptive behaviours). Maintaining NCI, which is necessary if the theorist wants to make welfare evaluation, implies rejecting CS and recognizing 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), when we consider the counterfactual enlightened choices of the individual as the ‘correct’ preferences (behavioural paternalism), or even 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 point is that identifying a way to infer welfare from inconsistent choices crucially depends on the *values* that are deemed important for conducting welfare analysis. This aspect has, however, 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 particular characterization of individual welfare. We have proposed a value of Consumer Autonomy (CA) as a weaker form of Consumer Sovereignty (CS),

which requires a certain degree of knowledge by the individual to ensure he makes autonomous choices.

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

**Acknowledgements.** 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 acknowledges the financial support of the FAST project (Facilitate public Action to exist from peSTicides) conducted by the Agence Nationale de la Recherche (ANR), reference 20-PCPA-0005. There are no competing interests. We are grateful to two anonymous referees, Franz Dietrich, Marc Fleurbaey, Glenn Harrison, Don Ross, and Bele Wollesen for their valuable feedback. We also thank the participants of the 2023 International Network for Economic Method conference in Venice for their comments. Any errors remain our own.

## References

- Aivazian V.A., J.L. Callen and I. Lipnowski 1987. The Coase theorem and coalitional stability. *Economica* 54, 517–520.
- Akerlof G.A. and R.J. Shiller 2015. *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton: Princeton University Press.
- Apestequia J. and M.A. Ballester 2015. A measure of rationality and welfare. *Journal of Political Economy* 123, 1278–1310.
- Arkes H.R., G. Gigerenzer and R. Hertwig 2016. How bad is incoherence? *Decision* 3(1), 20–39.
- Arrow K.J. 1951 [2012]. *Social Choice and Individual Values*, 3rd ed. New Haven: Yale University Press.
- Atkinson A. and J. Stiglitz 2015. *Lectures on Public Economics*, updated ed. Princeton: Princeton University Press.
- Bacharach M. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton: 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, 12–68.
- Bernheim B.D. and A. Rangel 2007. Toward choice-theoretic foundations for behavioral welfare economics. *American Economic Review* 97, 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, 51–104.
- Binder C. 2014. Plural identities and preference formation. *Social Choice and Welfare* 42, 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, 1211–1254.
- Chambers C.P. and T. Hayashi 2012. Choice and individual welfare. *Journal of Economic Theory* 147, 1818–1849.
- Chetty R. 2015. Behavioral economics and public policy: a pragmatic perspective. *American Economic Review* 105(5), 1–33.
- Christman J. 2009. *The Politics of Persons: Individual Autonomy and Socio-historical Selves*. Cambridge: Cambridge University Press.
- 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, 560–580.
- Dalton P.S. and S. Ghosal 2012. Decisions with endogenous frames. *Social Choice and Welfare* 38, 585–600.
- DellaVigna S. 2009. Psychology and economics: evidence from the field. *Journal of Economic Literature* 47, 315–372.
- Dennett D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.
- Dietrich F. and C. List 2013a. A reason-based theory of rational choice. *Noûs* 47, 104–134.
- Dietrich F. and C. List 2013b. Where do preferences come from? *International Journal of Game Theory* 42, 613–637.

- Dietrich F. and C. List** 2016. Reason-based choice and context-dependence: an explanatory framework. *Economics and Philosophy* **32**, 175–229.
- Echenique F., T. Imai and K. Saito** 2023. Approximate expected utility rationalization. *Journal of the European Economic Association* **21**, 1821–1864.
- Gao X.S., G.W. Harrison and R. Tchernis** 2023. Behavioral welfare economics and risk preferences: a Bayesian approach. *Experimental Economics* **26**, 273–303.
- Gold N.** 2022. Guard against temptation: intrapersonal team reasoning and the role of intentions in exercising willpower. *Noûs* **56**, 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**, 347–365.
- Graaff J.d.V.** 1963. *Theoretical Welfare Economics*, reprinted ed. Cambridge: Cambridge University Press.
- Harrison G.W. and J.M. Ng** 2016. Evaluating the expected welfare gain from insurance. *Journal of Risk and Insurance* **83**, 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**, 42–67.
- Harrison G.W. and D. Ross, eds.** 2023. Behavioral welfare economics and the quantitative intentional stance. In *Models of Risk Preferences: Descriptive and Normative Challenges*. Emerald Press.
- Harrison G.W., K. Morsink and M. Schneider** 2020. Do no harm? The welfare consequences of behavioural interventions. Technical report, CEAR Working Paper 2020.
- Hédoïn C.** 2015. From utilitarianism to paternalism: when behavioral economics meets moral philosophy. *Revue de Philosophie Économique* **16**, 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.** 1936. *Economists and the Public: A Study of Competition and Opinion*. J. Cape.
- Hutt W.H.** 1940. The concept of consumers' sovereignty. *The Economic Journal* **50**, 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**, 1–25.
- Kahneman D.** 1999. Objective happiness. In *Well-being: The Foundations of Hedonic Psychology*, eds D. Kahneman, E. Diener and N. Schwarz, 3–25. New York: Russell Sage Foundation.
- Kahneman D., B.L. Fredrickson, C.A. Schreiber and D.A. Redelmeier** 1993. When more pain is preferred to less: adding a better end. *Psychological Science* **4**, 401–405.
- Kahneman D., P.P. Wakker and R. Sarin** 1997. Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics* **112**, 375–406.
- Lecouteux G.** 2021a. Behavioral welfare economics and consumer sovereignty. In *The Routledge Handbook of Philosophy of Economics*, 56–66. London: Routledge.
- Lecouteux G.** 2021b. Welfare economics in large worlds: welfare and public policies in an uncertain environment. In *Elgar Modern Guide to the Philosophy of Economics*, eds H. Kincaid and D. Ross. Cheltenham: Edward Elgar.
- Lecouteux G.** 2022. Reconciling normative and behavioural economics: the problem that cannot be solved. In *The Positive and Normative in Economic Thought*, eds S. Badié and A. Grivaux, 148–166. London: Routledge.
- Lecouteux G.** 2023. The homer economicus narrative: from cognitive psychology to individual public policies. *Journal of Economic Methodology* **30**, 176–187.
- Lecouteux G. and I. Mitrouchev** 2024. The view from Manywhere: normative economics with context-dependent preferences. *Economics and Philosophy* **40**, 374–396.
- Lerner A.P.** 1972. The economics and politics of consumer sovereignty. *The American Economic Review* **62**, 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**, 1795–1810.
- Manzini P. and M. Mariotti** 2014. Welfare economics and bounded rationality: the case for model-based approaches. *Journal of Economic Methodology* **21**, 343–360.



- Mas-Colell A., M.D. Whinston and J.R. Green 1995. *Microeconomic Theory*. Oxford: Oxford University Press.
- McQuillin B. and R. Sugden 2012. Reconciling normative and behavioural economics: the problems to be solved. *Social Choice and Welfare* 38, 553–567.
- Mill J.S. 1849. *Principles of Political Economy*, 2nd ed. London: John W. Parker, West Strand.
- Mill J.S. 1859. *On Liberty*. Broadview Press.
- Mitrouchev I. 2019. Normative economics without the concept of preference. *OEconomia. History, Methodology, Philosophy* 9, 135–147.
- Mitrouchev I. 2024. Normative and behavioural economics: a historical and methodological review. *European Journal of the History of Economic Thought* 31(4), 533–562.
- Mongin P. 2006. Value judgments and value neutrality in economics. *Economica* 73(290) 257–286.
- Redelmeier D.A. and D. Kahneman 1996. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain* 66(1), 3–8.
- Rizzo M.J. and D.G. Whitman 2009. The knowledge problem of new paternalism. *BYU Law Review* 2009, 905–968.
- Ross D. 2005. *Economic Theory and Cognitive Science: Microexplanation*. Cambridge, MA: MIT Press.
- Rubinstein A. and Y. Salant 2012. Eliciting welfare preferences from behavioural data sets. *The Review of Economic Studies* 79, 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*. Chichester: John Wiley and Sons.
- Sen A. 1970. The impossibility of a Paretian liberal. *Journal of Political Economy* 78, 152–157.
- Sen A. 2017. *Collective Choice and Social Welfare*, expanded ed. London: Penguin Books.
- Steedman I. and U. Krause 1986. Goethe's Faust, Arrow's possibility theorem and the individual decision-taker. In *The Multiple Self*, ed. J. Elster, 197–231. Cambridge: Cambridge University Press.
- Sugden R. 1985. Liberty, preference, and choice. *Economics & Philosophy* 1, 213–229.
- Sugden R. 2004. The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. *American Economic Review* 94, 1014–1033.
- Sugden R. 2018a. *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: 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, 1–8.
- Thaler R.H. and C.R. Sunstein 2003. Libertarian paternalism. *American Economic Review* 93, 175–179.
- Thaler R.H. and C.R. Sunstein 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*, revised and expanded ed. London: Penguin Books.
- Thoma J. 2021. On the possibility of an anti-paternalist behavioural welfare economics. *Journal of Economic Methodology* 28, 350–363.
- Tversky A. and D. Kahneman 1981. The framing of decisions and the psychology of choice. *Science* 211, 453–458.
- Tversky A. and I. Simonson 1993. Context-dependent preferences. *Management Science* 39(10), 1179–1189.
- Varian H.R. 1987 [2014]. *Intermediate Microeconomics: A Modern Approach*, 9th ed. New York: W.W. Norton & Company.

## Appendix A: Proofs

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

*Proof that CCI and NI imply CS.* By CCI we know that there are not two contexts  $\gamma$  and  $\gamma'$  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  $\gamma$  and  $\gamma'$  for  $C$  and  $\succ$ , which means that  $\succ_{\gamma} = C_{\gamma'}$ .

**Guilhem Lecouteux** is an Associate Professor of economics at Université Côte d'Azur in Nice, France. His research focuses on the intersection of behavioural economics, the history of economic thought, and philosophy, exploring the role of behavioural sciences in the design and justification of public policies. Email: [guilhem.lecouteux@univ-cotedazur.fr](mailto:guilhem.lecouteux@univ-cotedazur.fr) URL: <https://sites.google.com/site/guilhemlecouteux/>.

**Ivan Mitrouchev** is a postdoctoral researcher in experimental economics at the French National Research Institute for Agriculture, Food and Environment (INRAE) in Grenoble, France. His research interests include addressing the philosophical challenges of evaluating welfare when individuals deviate from rationality principles. URL: <https://www.ivanmitrouchev.com/>.