A Brief History of Normativity in Behavioural Economics

Ivan Mitrouchev. Univ. Grenoble Alpes, INRAE, CNRS, Grenoble INP, GAEL, 38000 Grenoble, France. ivan.mitrouchev@inrae.fr

Malte Dold. Pomona College. Economics Department. 425 N. College Avenue. Claremont, CA 91711. malte.dold@pomona.edu

This version: 27 March 2025

Abstract. Behavioural economics is best known for explaining how people actually make decisions. However, less attention has been given to its role in shaping *normative* decision making, which focuses on evaluating welfare and guiding choices based on real-world decision patterns. This article traces the multiple meanings of normativity in behavioural economics from the 1970s to the 2010s, exploring different interpretations of "irrational" behaviour and their implications for policymaking.

Keywords. *logic – mistake – policy – rationality – welfare*

JEL codes. B21 – B41 – D90 – I31

Acknowledgements. 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. A preliminary draft was presented at the 2021 International Network for Economic Method (INEM) conference. We thank the audience for comments. We also thank Niels Boissonet, Jean-Sébastien Gharbi, Wade Hands, Dorian Jullien, Yao Thibaut Kpegli, Jérôme Lallement, Ramzi Mabsout and Robert Sugden for helpful comments on early versions. All mistakes remain ours.

1. Introduction

Behavioural economics started as a descriptive, explanatory and predictive enterprise. In a series of influential contributions (Tversky and Kahneman 1973, 1974, 1981, 1986; Kahneman and Tversky 1979), the heuristics-and-biases programme sought to (i) explore how heuristics lead to errors of judgement over objective probability, (ii) collect empirical findings that individuals systematically deviate from principles of rational choice, and (iii) propose a novel theoretical framework to describe, explain, and predict actual choice.[1] Although the early experiments of Daniel Kahneman and Amos Tversky revealed systematic deviations from expected utility theory, the normative benchmark of their research programme was still *rational choice* in the form of rules for decision making such as Bayesian updating, logic, and the axioms of expected utility theory. At this time, there was no particular focus on the evaluation, recommendation and prescription of policies based on their findings. But from the 1990s onwards, influential contributors to behavioural research—among them Daniel Kahneman and Richard Thaler—begun to direct a consequent part of their research to welfare evaluation, as well as the recommendation and prescription of policy.

In this article we provide a brief history of normativity in behavioural economics: how it started, how it evolved with different interpretations of "irrational" behaviour, and how it led to real-world policymaking.[2] Our analysis begins with the heuristics-and-biases programme in the 1970s-1980s, which centred on the rationality principles of decision making (Section 2). Then, in the 1990s, Kahneman and colleagues proposed a normative theory based on behavioural insights (Section 3). During the 2000s, varying interpretations of behavioural insights gave rise to distinct normative frameworks, each resulting in markedly different policy recommendations. Among the most widely discussed is the "preference purification" approach that led to the policy tool of nudging (Section 4). In the 2010s, a new wave of approaches emerged in response to critiques of nudging and its methodological foundations. Those emphasised *agency* (self-determined choice) as the normative benchmark for policy recommendations (Section 5).[3] We conclude with some challenges and open questions of this literature (Section 6).

Before proceeding, it is important to clarify our use of "normativity" throughout this article. In the humanities and social sciences, normativity is a broad and multifaceted concept (Spohn 2021). It encompasses the idea that certain actions or beliefs are preferable to others and that specific rules or standards ought to be followed. While descriptive models aim to describe, explain, or predict behaviour, normative models set standards for "good" decision making and evaluate situations based on those standards. Occasionally, decision theorists and economists introduce a third category: the prescriptive (Bell et al. 1988). We consider the prescriptive as a practical extension

[1] The origins of behavioural economics can be traced back to Herbert Simon's work in the 1950s (Simon 1955, 1956). However, we begin our historical analysis with the heuristics-and-biases programme, as it has become the dominant framework in contemporary behavioural economics. See Heukelom (2014).

[2] For other histories, which focus on different aspects of behavioural economics, see Lecouteux (2016), Moscati (2018) and Viale (2022). See also Heukelom (2014: Ch. 4), who provides a discussion of how the descriptive/prescriptive relationship in behavioural economics stabilised, and how the normative role of rational choice theory evolved through the various stages of Daniel Kahneman and Amos Tversky's research.

[3] In contrast to the rest of the sections, Section 5 is more methodological/philosophical than historical, given that it primarily engages with recent contributions.

of the normative, applied to real-world individuals, and therefore to policymaking. We treat the prescriptive as a subset of the normative and highlight instances where the prescriptive emerges within the approaches we discuss.

2. 1970s-1980s: Rational Choice as The Gold Standard

One of the earliest tensions between rational decision-making models and observed human behaviour can be traced back to the famous Allais (1953) paradox, which challenged the foundational assumptions of expected utility theory. Notably, Leonard Savage himself violated the *independence of irrelevant alternatives* axiom of expected utility theory when presented with Allais' hypothetical choice task. Although Savage acknowledged that the paradox exposed a conflict between his theoretical framework and his intuitive preferences, he maintained that expected utility theory was fundamentally correct as a *normative* model of decision making. He recognised that his intuitions led to inconsistent choices, but argued these were instances of human error rather than flaws in the theory itself. Because Savage considered the *independence* axiom desirable to follow in virtue of its logical structure, he ultimately chose to align with it, suggesting that when individuals, including himself, make decisions that are inconsistent with this principle, they should reconsider their choices. Although behavioural economics had not yet been recognised as a distinct field, this event marked an early instance of how the discipline would later interpret deviations from decision-making rules—specifically as *errors* in judgement, or, simply put, *mistakes*.[4]

The heuristics-and-biases programme was developed primarily by Kahneman and Tversky in the 1970s. The aim was to explore how people make decisions and judgments under uncertainty. The programme highlighted the relationship between heuristics and biases through lab experiments where participants were asked to solve cognitive choice tasks. Heuristics are mental shortcuts people use to make decisions or solve problems under various constraints—typically temporal (quick decisions), informational (complex or incomplete information), and cognitive (limited processing capacity). Due to these constraints, Kahneman and Tversky's experimental results showed that people's heuristics can lead to *biases*. Those are defined as systematic and predictable judgments that arise from reliance on heuristics. This does not mean proponents of the heuristics-and-biases programme consider a bias to be an error (or mistake) in itself, but more precisely, when it deviates *from a given benchmark*.[5]

[4] From a theoretical viewpoint, this was the beginning of a sharp break from the mid-20th-century Walrasian-Paretian welfare theory, which relied on widely accepted assumptions. Individuals were seen as having stable preferences and maximising them, with welfare defined by preference satisfaction. The perfectly competitive market served as the institutional baseline, where exchanges occurred at equilibrium prices and markets always cleared. Social welfare was largely tied to Pareto efficiency, with the first fundamental theorem of welfare economics stating that all general equilibria were Pareto efficient. In this model, individuals made no mistakes, as their choices in competitive markets maximised welfare. Because our historical analysis starts from the advent of the heuristics-and-biases programme in the 1970s, we bracket out these considerations and direct our reader to Hands (2024).

[5] A fundamental disagreement exists between proponents of the heuristics-and-biases programme and those of the fast-and-frugal heuristics programme regarding both the descriptive and normative interpretations of rationality. Todd and Gigerenzer (2012) do not view heuristics or deviations from standard rationality principles as biases. As Gigerenzer (1996: 102) puts it, "biases are not biases". Instead, the fast-and-frugal heuristics programme argues that certain heuristics lead to "good enough" decisions, with their effectiveness depending on the specific environment in which they are used. This follows Simon's (1956) satisficing criterion for evaluating a "good" decision, in contrast to the standard

This means that a bias is interpreted as a mistake when participants fail to provide a correct answer to a problem that is subject to objective and verifiable scrutiny. This interpretation was largely shared by Kahneman and Tversky at the time, who, in their proposition of the first generation of prospect theory, provided a general note about (supposed) self-acknowledged errors of reasoning by the decision maker who violates the axioms of expected utility theory:

> "These departures from expected utility theory must lead to normatively unacceptable consequences, such as inconsistencies, intransitivities, and violations of dominance. Such anomalies of preference are normally corrected by the decision maker when he realizes that his preferences are inconsistent, intransitive, or inadmissible. In many situations, however, the decision maker does not have the opportunity to discover that his preferences could violate decision rules that he wishes to obey. In these circumstances the anomalies implied by prospect theory are expected to occur." (Kahneman and Tversky 1979: 277)

Then Slovic and Tversky (1974) explicitly highlighted the foundational role of logic in establishing the desirability of rationality axioms:

> "Many decision theorists believe that the axioms of rational choice are similar to the principles of logic in the sense that no reasonable person who understands them would wish to violate them." (368)

The idea that logic instructs us how we ought or ought not to think or reason is known as the *normative status of logic* (Steinberger 2017). That is, we consider it to be "a bad thing" to be inconsistent over logical principles, and conversely, we consider it to be "a good thing" to be consistent over logical principles. Because many rationality principles—such as *transitivity* and *independence*—are grounded in logical principles, decision theorists at the time regarded logic as the normative foundation of rational choice. Rationality principles and logic principles both centre on the idea of consistency, but in different domains. Logic is concerned with consistency in propositional content, where contradictions like "p ∧ ¬p" violate the basic rules of reasoning. Rational choice theory, by contrast, focuses on consistency in preference relations—how individuals rank different outcomes.

The first major challenge in interpreting normativity within the behavioural paradigm along those lines stemmed from the nature of the rationality principles under discussion. Since not all rationality principles in the decision-making models under consideration were grounded in formal logic, the criteria for determining the desirability of such principles became ambiguous. If not grounded in logic, what determines the normative appeal of these principles? As an illustration, consider the three heuristics identified by Kahneman and Tversky in the 1970s as follows. The first is *availability* (Tversky and Kahneman 1973, 1974), which is the tendency to estimate the likelihood of an event based on how easily instances of that event come to mind (for example, people typically overestimate the risk of aeroplane crashes after hearing about a crash on the news). The second is *representativeness* (Tversky and Kahneman 1974), which is about judging the probability of an event based on how similar it seems to a stereotype (for example, when people know someone to be quiet and methodical, they

maximisation criterion. More fundamentally, the disagreement between the two programmes stems from conflicting interpretations of probability—specifically, Bayesian versus frequentist interpretations. For an overview of this debate, see Gigerenzer (1991, 1996) and Kahneman and Tversky (1996).

often think that person is more likely to be a librarian than a salesperson). The third is *anchoring* (Tversky and Kahneman 1974), i.e. the tendency to rely heavily on the first piece of information when forming judgments and making decisions (for example, when people negotiate a price, the initial offer strongly influences the final deal).

Availability and *representativeness* are heuristics in uncertainty that lead to probability distortions, thereby violating Bayes' theorem. For a Bayesian, any violation of Bayes' theorem is seen as illogical and thus normatively unappealing, given the normative status of logic. *Anchoring*, however, differs from these heuristics as it is not inherently tied to uncertainty and does not result in a violation of any logical principle. However, this distinction was rarely discussed among behavioural economists at the time. Instead, behavioural economists tended to treat "rational thinking" as a general normative benchmark, without specifying which rationality principles are at stakes, and (perhaps most importantly) *in virtue of what underlying principles* it was desirable to follow those.

The same problem applies for the "framing effect", identified by Tversky and Kahneman (1981). *Framing* is defined as the tendency to make different decisions or judgments based on how information is presented, rather than considering the actual content of the information itself. Like anchoring, framing does not inherently relate to logic either. The only connection arises when a rationality principle is formalised—such as in expected utility theory, where the *invariance* principle requires that different representations of a choice problem yield consistent preferences. Unlike formal principles, framing relies on the specific language used and psychological or social cues influencing individuals' judgement. As Tversky and Kahneman (1981) noted,

> "The frame that a decision-maker adopts is controlled partly by the formulation of the problem and partly by the norms, habits, and personal characteristics of the decision-maker." (453)

Some biases lead to distorted perceptions regarding what is *objectively* true (what is subject to verifiable scrutiny). For example, the *overconfidence bias*—the tendency to overestimate one's own abilities or the accuracy of one's judgments—and the *confirmation bias*—the tendency to seek out or interpret information in a way that confirms pre-existing beliefs or hypotheses—fall into this category. But this is not the case for some other biases like *framing*, which are rather subject to *subjective* evaluation. Perhaps the most famous example is *loss aversion*, which is the tendency to prefer avoiding losses to acquiring equivalent gains. Here, it is up to the individual to determine whether acting according to loss aversion constitutes a mistake, as there is no external objective benchmark—unlike the rules of logic—against which loss aversion can be classified as a "bias".

In a nutshell, the ambiguity in interpreting normativity arose when it became difficult to categorise all normative principles under the same nature, as some were more closely aligned than others with the principles of logic and probability theory.

Following the "empirical" trend of questioning the descriptive validity of expected utility theory, a parallel line of research during this period was undertaken by Keneth MacCrimmon (1968), Herbert Moskowitz (1974), as well as Paul Slovic and Amos

Tversky (1974).[6] This group proposed experiments whose aim was to test the *normative appeal* of the rationality principles which were empirically violated. In this innovative line of research, preferences over axioms were elicited in specific choice problems. Subjects could revise their choices in various risk tasks, revealing if their adjustments aligned more closely with rational principles.[7]

The first study on choice revision by MacCrimmon (1968) provided choice tasks to business executives, allowing them to reconsider their responses after exposure to arguments for and against principles like *transitivity*, *independence of irrelevant alternatives*, and *dominance*. Although some participants admitted having made "mistakes", the author notes that these were often attributed to laziness or difficulty with the questions. Moskowitz (1974) explored the normative appeal of the *independence of irrelevant alternatives* principle through Allais-and-Morlat-type problems, presenting tasks in word, tree, and matrix formats. Participants initially made choices, then viewed other students' responses to judge logical coherence, and could discuss choices in one treatment group.[8] After discussions, participants were more likely to align with rational principles. Those without discussion opportunities showed little change in their responses. Across frames and treatments, most participants tended to follow rationality principles.[9] The experiment of Slovic and Tversky (1974) regarding the *independence of irrelevant alternatives* axiom, however, showed contrasting results. In their experiment, students were given choices in Allais and Ellsberg problems without social or corrective feedback, but with a clear understanding of the principles. Despite these conditions, the majority of participants maintained their choices even when these conflicted with the axiom.[10]

Given that these experiments showed mixed results in approval rates, it is unsurprising that divergent opinions arose about what constitutes "goodness" in the context of decision theory, with even the normative status of logic itself coming under scrutiny. It became evident that differing value judgments about what constitutes "goodness" were unavoidable, an important insight emphasised by MacCrimmon (1968), who proposed a pragmatic response to this issue:

> "A descriptive theory can be judged by its explanatory or predictive ability. It is more difficult, though, to judge a normative theory. Presumably, adopting a good normative

[6] See in particular Mongin (2019). These experiments were primarily motivated by Allais' (1953) paradox.

[7] This research on *choice revision* has seen recent updates with incentivised risk-based choice tasks, as explored by Benjamin et al. (2020), Nielsen and Rehbeck (2022), Breig and Feldman (2024) in choice under risk, and Andersson et al. (2023) in the domain of social redistribution. Surprisingly, however, this line of research largely halted at the end of the 1970s and lay dormant for nearly half a century. We briefly revisit this point in the concluding remarks.

[8] Note again the normative status of logic by the experimentalist, who considers *logic*, and not something else (e.g. welfare or happiness), as the normative benchmark.

[9] The lowest rate of approval was observed among subjects in the non-discussion group after receiving feedback, where only 50% chose to align with the *independence of irrelevant alternatives* principle.

[10] Only one experimental setting showed that a majority of participants (61%) preferred to follow the axiom, while in the three other settings, a majority chose to persist in rejecting it (59%, 66%, and 80%). Slovic and Tversky (1974) specifically designed these experiments in response to concerns that MacCrimmon's (1968) study may have been influenced by the experimenter demand effect and due to the peculiar characteristics of the sample. In their words: "subtle pressures, in combination with the cooperativeness of subjects participating in a training course for a prestigious job, may have influenced the subjects to conform to the axioms." (369).

> theory will lead to 'better' results. But 'better' in what sense? The criteria must be specified and will often be part of the theory itself. One condition we might expect a good normative theory to satisfy is that it should seem reasonable to individuals with expertise in the domain of usage. Thus, we should expect a good normative theory of decision to seem reasonable to successful, practising decision-makers." (3-4)

MacCrimmon (1968) recognised that, in the absence of clear criteria for what should be considered "good," it might be more appropriate to let the experts (such as the business executives in his study) determine what makes them "better off". This highlighted a trade-off in using behavioural insights for normative analysis. If the "good" is not defined *a priori* by an external observer (such as a theorist, economist, or policymaker), what is "best" in a choice task remains ambiguous, as experts may hold differing views. Conversely, if the "good" is predefined by the observer, the prescribed "best" choice could conflict with individuals' own assessments of what is best for themselves.

Thus, the challenge of establishing a clear normative benchmark in behavioural economics to differentiate between good and bad choices remained unresolved. From the 1990s onwards, Kahneman and colleagues sought to address this issue by exploring more expansive notions of welfare.

3. 1990s: The First Step Towards Welfare

The 1990s marked a significant shift away from examining the normative status of logical principles of rationality. A new research programme, led by Kahneman, focused more explicitly on understanding what it means for individuals to engage in welfare-improving choices. As noted earlier, behavioural insights had challenged the idea that observed preferences are a reliable indicator of welfare. This raised a fundamental question. If choice is not a good proxy for welfare, what else could it be? Could it be that people deviate not only from logical principles but also from welfare-maximising choices, particularly when welfare is defined in the traditional Benthamite terms of pleasure and pain? (Kahneman et al. 1997). The distinction between *decision utility* (what people choose) and *experienced utility* (what they feel in terms of pain/pleasure) had already been broached by Kahneman and Tversky (1984: 349-350), making the time ripe for further exploration of these ideas.

The first published experiment on this topic was conducted by Kahneman and Snell (1990), who presented evidence that individuals have difficulty accurately predicting their future experienced utility. Building on March's (1978) proposition that "decision utility" may diverge from "experienced utility," their study provided the first empirical support for that distinction. Kahneman and Varey (1991) then questioned the validity of using choice as the sole measure of utility, proposing that experienced utility comprises three elements: the experience itself, its memory, and its anticipation. Kahneman and Snell (1992) further explored whether people can predict their future hedonic experiences and concluded they often cannot. In an eight-day study where participants consumed ice cream while listening to music, they found a near-zero correlation between actual and predicted enjoyment ratings. While, according to the authors, these findings did not conclusively show an inability to predict future tastes, the authors interpreted it as *errors* in such predictions. In this sense, the "error of reasoning" perspective was directly applied to an approach where a clear-cut criterion—namely, the optimisation of hedonic states—could distinguish good choices from bad ones.

The question of establishing normative rules for aggregating hedonic states was then actively explored. Kahneman and Snell (1992) proposed three conditions for utility integration: *monotonicity* (adding pain should increase overall disutility), *non-discrimination* (two equivalent pain moments should equally contribute to overall disutility), and *additivity* (the increase in disutility from one pain point to the next should match the added experience's disutility). This was the first time that *axioms over welfare*—instead of axioms over *logical principles*—were proposed in the behavioural economic literature. In Kahneman et al.'s experiments, subjects endured discomforts like carrying a suitcase, putting one's hands in ice-cold water, sitting in a vibrating room, or standing uncomfortably. The results showed most participants violated these conditions. Notably, they found that adding pain could sometimes lower overall negative evaluation, contradicting *monotonicity*. These deviations, still viewed as errors, were then further investigated by Kahneman et al. (1993) and Fredrickson and Kahneman (1993). Taken together, these studies underscored that individuals often judge experiences more by their peaks and ends than by their cumulative duration, challenging traditional assumptions of utility theory and raising new questions about how we assess well-being based on memory.

Before the influential manifestos of asymmetric and libertarian paternalism (Camerer et al. 2003; Thaler and Sunstein 2003), Kahneman (1994) introduced the idea of welfare-improving paternalistic interventions motivated by behavioural insights. In particular, he suggested that individuals often struggle to accurately predict their future happiness, while experts might have superior knowledge about what choices effectively increase individuals' happiness and can thus guide their choices. At the end of the 1990s, this programme led by Kahneman aimed at constructing a comprehensive theory of "objective" happiness rooted in hedonic states experienced subjectively. The theory was based on experiments conducted by Kahneman and his co-authors in the 1990s, which yielded three general insights: (i) individuals often exhibit myopia in decision making, (ii) they frequently mispredict their future preferences, and (iii) their choices are shaped by fallible memories and distorted evaluations of past experiences. These findings led Kahneman (1999) to advocate for a broader definition of welfare, incorporating what he termed the "substantive" criterion of experienced utility. This criterion evaluates the relative "goodness" of decision outcomes based on people's actual hedonic experiences independently of individuals' observed preferences, representing a significant departure from traditional welfare economics, which primarily assesses welfare based on the satisfaction of individual preferences.

A theory for measuring experienced utility was thus needed to establish the conditions under which utilities could be meaningfully integrated over time. This effort was pioneered by Kahneman et al. (1997) in their famous "Back to Bentham" contribution, in which the authors proposed a normative theory for aggregating the temporal profiles of utility that individuals experience instantaneously. Following the utilitarian tradition, they proposed that a social planner could potentially maximise the sum of the total utility for each individual within an objective function. This marked a significant shift in *normativity* within behavioural economics. While the normative framework of the 1970s and 1980s was rooted in *logical* principles, Kahneman et al. introduced an *ethical* perspective of the good life, establishing a strict criterion for defining what constitutes happiness and how to measure it.

Unsurprisingly, this approach to normativity was recognised to have its own challenges and shortcomings (Kahneman and Sugden 2005). A key concern was that the idea of the good life could not be reduced to simple hedonic metrics. People might view the good life more as an accumulation of meaningful memories than as a continuous stream of pleasure and pain. Most fundamentally, using experienced utility as a policy guide may conflict with a core principle of liberal philosophy: the preservation of individual autonomy. This tension called for an alternative normative standard—one that upheld autonomy while recognising that individuals sometimes make choices that jeopardise their future welfare. The concept of "asymmetric" or "libertarian paternalism" then emerged as a potential way to reconcile these concerns.

4. 2000s: The Dominance of Libertarian Paternalism and *Nudge*

Following the growing interest towards welfare initiated by Kahneman and co-authors, the main challenge of the 2000s was as follows. If observed preferences were error-prone, and welfare grounded in hedonic states seemed insufficient, what alternative normative standard could serve as the benchmark for guiding policy analysis?[11] Two prominent approaches emerged to tackle this problem (Dold 2023; Fumagalli 2024). The first, known as the *preference purification approach*, aligns closely with the traditional framework in normative economics. The second, the *opportunity approach*, seeks to provide a framework for normative analysis without referring to individuals' preferences. These two approaches stem from distinct historical traditions regarding the role of normative economics (Mitrouchev 2024). In this article, we exclusively focus on preference purification, as it is the mainstream theoretical approach on which libertarian paternalism and nudging are grounded.[12]

The preference purification approach begins with the assumption that people's observed choices are often the result of "bad" judgements, i.e. judgements "they would not have made if they had paid full attention and possessed complete information, unlimited cognitive abilities, and complete self-control" (Thaler and Sunstein 2008: 5). Such a statement suggests that the normative standard relies on the notion of "undistorted" or "true" preferences (Sugden 2018: 65).[13] Some suggested that this perspective envisions an "inner rational agent" aligned with neoclassical theory, trapped within a bias-prone "psychological shell" (Infante et al. 2016). According to this view, individuals aim to act on a core set of well-integrated preferences that remain

[11] Although this question marked an original contribution to integrating behavioural insights into welfare analysis at the time, the underlying idea was far from new. In particular, Harsanyi (1977: 646), as a defender of utilitarianism, coined the term "manifest preferences" to describe an individual's "actual preferences as manifested by his observed behaviour, including preferences possibly based on erroneous factual beliefs, or on careless logical analysis, or on strong emotions that at the moment greatly hinder rational choice."

[12] For a discussion of the opportunity approach, see Schubert (2015), Mitrouchev (2019), and Dold and Rizzo (2021).

[13] As this article aims to provide a brief history of normativity in behavioural economics, we do not delve into all the nuances of the preference purification approach. Notably, we do not address generalisations, such as those proposed by Bernheim (2016, 2021) that are beyond the assumption of true preferences. For a discussion of the historical roots of the preference purification approach, see Hands (2024). For a comprehensive overview of approaches employing the preference purification approach, see Bernheim (2016) and Sugden (2018: Ch. 4).

consistent, but psychological biases often derail them during the decision-making process.[14] Favoured by many behavioural economists, this approach aligns closely with standard welfare economics, relying on the key assumption that people have underlying "true" preferences. As Sugden (2018: 62–63) argues, without this assumption, the approach would lack a clear normative standard for evaluating preferences. In practice, the logic of the preference purification approach has led behavioural economists to develop extensive lists of "biases" to explain deviations between observed choices and welfare-improving choices (Rizzo and Whitman 2020). Additionally, this approach has contributed to designing *prescriptive policies* that leverage behavioural insights with the goal to help individuals satisfy their "true" preferences. One of the most widely discussed applications of this strategy is the implementation of *nudges*, defined as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives." (Thaler and Sunstein 2008: 6).

Nudges, following the logic of libertarian paternalism (Thaler and Sunstein 2003), aim to steer behaviour from "faulty" preferences towards "true" preferences (the paternalistic aspect) without restricting options or significantly altering incentives (the libertarian aspect). Ultimately, they are designed to "[make] choosers better off, as judged by themselves" (Thaler and Sunstein 2008: 5). However, it is not always clear what the "as judged by themselves" clause means in policy practice. In fact, libertarian paternalism often grapples with the conflation between rational thinking—defined as "unlimited information, ability, and self-control"—and what individuals perceive as beneficial for themselves. Due to this conflation, a serious challenge arises. The concepts of unlimited information and ability are determined by external observers, such as economists or policymakers, while the subjective assessment of what makes people better off remains inherently internal and inaccessible due to what Rizzo and Withman (2009) refer to the "knowledge problem".

This issue echoes the insights of MacCrimmon (1968), who identified a critical trade-off in the realm of *normativity* within behavioural economics. On the one hand, welfare analysis can be restricted to a narrow scope, relying on decision-making experts to establish criteria for rationality (e.g. *dominance* and *transitivity*). On the other hand, if we choose to generalise the normative approach to encompass a broader population, the criteria for what constitutes a "good" decision become ambiguous and less transparent. Libertarian paternalism provides a more flexible interpretation of "mistake" due to its unclear (and potentially shifting) normative benchmark. Nudges can be based on various mechanisms, such as social norms, expert judgments, the behaviour of experienced choosers, long-term goals, or the minimisation of opt-outs in default settings (status quo), among others. Across these mechanisms, the normative benchmark may differ—for instance, depending on which social norms are emphasised or at which point in time the goals of an experienced chooser are elicited. Although this posed some issues in the academic literature, leading to various discussions about the philosophical and methodological problems of libertarian paternalism and nudges (Rizzo and Whitman 2009; Grüne-Yanoff 2012; Hands 2020;

[14] As Sunstein (2018: 7) states: "[it] is psychologically fine to think that choosers have antecedent preferences, but that because of a lack of information or a behavioural bias, their choices will not satisfy them."

among many others), it did not delay the application of the preference purification approach in real-world policymaking.[15]

5. 2010s-today: Agency-Centric Approaches

Building on the international success of *Nudge* (Thaler and Sunstein, 2008) and the growing application of the preference purification approach, the field of *behavioural public policy* experienced significant expansion throughout the 2010s. This growth was marked by a surge in policy reports spanning domains like health, transport, and finance across numerous countries, including the United States, the United Kingdom, Australia, and various European nations (OECD 2017). The establishment of the *Behavioural Public Policy* journal in 2017 further underscored the expanding influence of behavioural insights in public policy. However, not all scholars welcomed the field's trajectory, particularly the dominance of nudging both in policy practice and academic discussion (Oliver 2023; Chater and Loewenstein 2024). Nudging focuses primarily on changing behavioural *outcomes*, sidestepping deeper discussions about what it means to be fully rational and "bias-free", and whether this is genuinely a welfare-relevant state for individuals (Rizzo and Whitman 2020).

In response to the limitations of the preference purification approach, a growing number of scholars have recently advocated for *agency* as the normative yardstick in behavioural public policy.[16] In this literature, agency is understood either more objectively as the capability to form reasoned intentions and act on them (Banerjee et al. 2024) or, more subjectively, as the decision-maker's sense of competence and autonomy (Dold et al. 2024).[17] Although these approaches differ in their conceptualisation of agency, they share a common critique of strategies used in behavioural public policy that (a) treat behavioural outcomes as target variables and (b) rely on exploiting citizens' cognitive biases to achieve those outcomes. In contrast to nudges, which often capitalise on such biases, agency-centric approaches focus on improving the quality of the cognitive processes leading to choice.

Proponents of agency-centric approaches critique existing behavioural public policy frameworks for how they define welfare. Some rely on a first-person standpoint, focusing only on individuals' current preferences while ignoring how context shapes decisions. Others take a third-person approach, overriding individual preferences in favour of theorists' views on what constitutes welfare. In contrast, agency-centric approaches advocate for a "second-person" standpoint that prioritises individuals' capacity to engage with and reflect on the various contexts shaping their choices (see Lecouteux and Mitrouchev 2024 for an attempt to theorise these approaches). In these approaches, there are certain conditions that are essential to evaluate welfare from the perspective of the individuals themselves. These conditions include (1) *appropriate cognitive abilities*, (2) an *adequate range of options*, and (3) *independence from*

[15] See Hands (2024) for an in-depth analysis of the issues related to the preference purification approach. Alternative approaches to preference purification have also been proposed (Sugden 2004, 2018). See Mitrouchev (2019, 2024) and Dold and Schubert (2018) for reviews.

[16] See, for instance, Banerjee et al. (2023), Dold and Lewis (2023), Grüne-Yanoff and Hertwig (2016), Hertwig and Grüne-Yanoff (2017), Hargreaves Heap (2013, 2017, 2023).

[17] In self-determination theory, autonomy refers to "a sense of initiative and ownership in one's actions," while competence entails "the feeling of mastery, a sense that one can succeed and grow" (Ryan and Deci 2020: 1).

manipulation.[18] A range of prescriptive interventions have been proposed to address these three conditions.

To address condition 1 ("cognitive abilities"), boosts have been suggested as an alternative to nudges (Grüne-Yanoff and Hertwig 2016; Hertwig and Grüne-Yanoff 2017). Boosts aim to enhance decision-makers' cognitive abilities to help them achieve their objectives "without making undue assumptions about what those objectives are" (Grüne-Yanoff & Hertwig 2016: 156). Unlike traditional consumer protection policies, which often aim to improve informational input, boosts focus on expanding agents' cognitive strategies to transform information into choices. The idea is to educate individuals on the effective use of decision heuristics—simple rules of thumb that are effective in specific environments and often outperform existing cognitive strategies. One example of a boost is teaching individuals how to convert one risk format into another, such as translating relative probabilities into natural frequencies (Hertwig and Grüne-Yanoff 2017: 977). Another example is teaching individuals to use fast-and-frugal decision trees (FFTs) as a diagnostic tool, enabling quick and effective diagnoses based on only a few informational cues (Marewski and Gigerenzer 2022).[19]

Unlike nudges, boosts require active cooperation from individuals, who must decide whether to engage with them or not. For boosts to be effective, individuals must accept the training, internalise the competence, and apply it when needed. These factors are supposed to ensure that behaviour changes resulting from boosts are grounded in reason. In addition to boosts that can help individuals achieve specific objectives more effectively (*instrumental reasoning*), some proposed agency-enhancing interventions are "educational" in nature, aiming to empower individuals to reflect on their evolving, context-dependent objectives (*substantive reasoning*). To enhance individuals' substantive reasoning capacities, agency-centric approaches to behavioural public policy focus on the conditions that foster reflection on which preferences to hold—such as the educational system, media, and the arts (Hargreaves Heap 2013). The effectiveness of these institutions can be assessed by their ability to empower individuals to choose life plans with which they identify themselves, ensuring they have the necessary resources to reflect on their preferences and act accordingly.

To address condition 2 ("adequate range of options"), agency-enhancing behavioural public policy stress the institutional foundations required for a dynamic society where individuals can explore social influences and develop through Millian "experiments in living". Such experiments expose people to a diverse range of perspectives and identities, providing the "raw material" (examples of different lifestyles and paths) to help them shape their own preferences and identities (Delmotte and Dold 2022). This perspective underscores the importance of behaviourally informed *s-frame interventions* that tackle structural barriers to such experimentation, potentially including income inequality and social obstacles to equal opportunities (Chater and

[18] These are the three "classic" conditions of autonomy outlined by Raz (1986) in *The Morality of Freedom*. In this article, we remain neutral on whether these conditions are necessary, sufficient, or neither.

[19] The boost literature has generated a large array of policy proposals. These policies can be broadly categorised into three types: (1) those that enhance risk competence in scenarios where risks are known and measurable, (2) those that build domain-specific competence by teaching effective behavioural heuristics, and (3) those that train individuals to use fast-and-frugal decision trees for navigating situations of uncertainty.

Loewenstein 2023).[20] Furthermore, laws and policies that safeguard core *civil liberties* (such as freedom of expression, freedom of movement, and freedom from discrimination) can enhance agency by enabling individuals to engage in "experiments in living". Empirical evidence shows that these freedoms positively and directly impact people's sense of autonomy (Ryan and DeHaan 2023). This institutional approach sharply contrasts with the nudging paradigm, which seeks to prevent errors through *i-frame interventions*, whereas learning through "experiments in living" inherently involves making and reflecting on mistakes as part of personal growth.

To address condition 3 ("independence from manipulation") and enable individuals to pursue their own version of a good life, Oliver (2018, 2022) advocates for regulatory interventions to mitigate behavioural-informed harms. Central to this proposal is the concept of "budges," a type of regulation aimed at addressing "behavioural externalities" in exchange relationships. Positioned as a middle ground between *laissez-faire* policies and overly paternalistic interventions, budges aim to ensure fairness and reduce manipulation in market transactions. Unlike nudges, which often exploit cognitive biases to steer behaviour, budges specifically target manipulative practices that undermine free and fair exchange. Oliver highlights how businesses leverage insights of behavioural economics, such as present bias and loss aversion, to manipulate consumers through tactics like misleading advertisements or complex pricing structures. Oliver argues that these practices justify regulatory intervention by causing substantive harm in exchanges (e.g. people overconsume certain goods and show severe post-consumption regret). Examples of effective budges include regulations on payday loans or misleading gambling advertisements, both of which exploit behavioural vulnerabilities to the detriment of consumers. Ultimately, within this approach, society needs to discuss and determine what constitutes undue harm (Oliver 2023).

The approaches discussed in this section, with their emphasis on the social conditions for individual agency, offer a shift in the debate on the normative implications of behavioural economics. They move away from the traditional i-frame focus on modifying individual behaviour through nudges and instead highlight the importance of systemic changes (the s-frame), such as regulations and institutional reforms. While i-frame interventions often rest on vague and problematic normative standards (as discussed earlier), they also tend to produce modest or negligible results, failing to drive meaningful societal change. By framing problems as individual shortcomings rather than structural issues, i-frame approaches risk diverting attention and resources from systemic solutions (Chater and Loewenstein 2023). Agency-centric approaches advocate for a balance between i-frame and s-frame strategies, with the latter taking precedence in addressing large-scale challenges to individual agency.

Yet a critical observation is warranted. While agency-centric approaches highlight the structural conditions necessary for fostering agency, they do not definitively establish what agency is. Any normative model of agency used in public policy is inherently a "thick concept", as it simultaneously describes and evaluates matters (Alexandrova

[20] According to Chater and Loewenstein (2023), i-frame analysis focuses on individual-level solutions to policy problems, assuming that adverse outcomes stem from human cognitive frailties (e.g. present bias, bounded willpower, etc.). S-frame analysis emphasises systemic changes, addressing the institutional and structural factors shaping individual choices (e.g. laws, norms, narratives). In a nutshell, they aim to "fix" the rules of the game rather than the players.

and Fabian 2022). The effectiveness of such models should be assessed based on whether individuals feel adequately represented by them. It remains an open question whether people resonate more with the "subjective" model, emphasising the sense of agency, the "objective" model, focusing on opportunities and reasoning capabilities, or whether agency is even a significant concern for individuals. By raising awareness of how social conditions shape beliefs and preferences or prompting individuals to reflect on their own decision-making errors, agency-centric approaches might be perceived as intrusive or unsettling.

Ultimately, these are empirical issues that can be resolved by actively involving affected citizens in co-creating policies. This aligns with Chater (2022: 1), who argues that behavioural insights "do not override, but can (among many other factors) inform, our collective decision-making process." The primary role of behavioural insights in public policy is to inform and enrich public debate when determining the rules by which we wish to live. Academic expertise can play a vital role in enhancing public deliberation by helping citizens and policymakers better understand the social conditions and challenges associated with individual agency. Admittedly, public deliberation is not without flaws, as it can exacerbate decision-making issues such as motivated reasoning, herd behaviour, and groupthink. Nevertheless, when guided by inclusive and well-designed rules of discourse, deliberative processes might be able to help citizens articulate and share their beliefs about agency-centric behavioural public policy (Colin-Jaeger and Dold 2025).

6. Concluding Remarks

In this article, we traced the shifting meanings of normativity in behavioural economics: from the emphasis on logic and rationality axioms in the heuristics-and-biases programme of the 1970s and 1980s, to the rise of experienced utility in the 1990s, to the focus on "true" preferences in the nudging approach of the 2000s, and, more recently, to the exploration of agency as an alternative benchmark. Based on this brief history, we aim to highlight two significant challenges that warrant further attention.

First, regarding the heuristics-and-biases programme, scholars have criticised the lack of empirical evidence demonstrating that deviations from rationality principles systematically leave individuals worse off (Gigerenzer 2018; Sugden 2019; Rizzo and Whitman 2020). One can see the revitalisation of the choice revision literature as a promising response to this critique. It examines how individuals update or revise their choices under risk (Benjamin et al. 2020; Nielsen and Rehbeck 2022; Breig and Feldman 2024), under certainty (Crosetto and Gaudel 2023) and under social redistribution (Andersson et al. 2023). A promising avenue of research in this experimental literature would be to include more qualitative reports, in a way that participants could clarify whether they revised their choices towards a more "rational" direction due to error correction, or for other reasons, such as changing their minds, uncertainty about their preferences, or a deliberate desire to diversify their choices. To our knowledge, this direction has (so far) not been undertaken.

Second, the growing emphasis on agency and citizen participation within behavioural public policy circles appears to conflict with the public's demand for paternalistic interventions. Meta-analyses provide suggestive evidence that people are generally willing to be nudged across various domains (Reisch and Sunstein 2016, among others). While these findings should be interpreted cautiously due to methodological differences in assessing individuals' willingness to be nudged, they challenge the assumption that citizens consistently prefer agency and participation over nudging and top-down interventions in their decision-making processes.

A potential lesson from our historical analysis is that, in the absence of better empirical evidence about what citizens consider "good" choosing to be, we (as economists) should exercise caution and refrain from being overly enthusiastic about using behavioural insights to, in Thaler's (2015: 307) words, "make the world a better place". With the rise of behavioural public policy, normativity has become increasingly fragmented, as behavioural interventions often assume normative benchmarks on an *ad hoc* basis (e.g. exponential discounting for intertemporal decisions like savings). These benchmarks seem to emerge "from nowhere" (Sugden 2018). Unlike earlier approaches that explicitly identified the axioms underpinning normative benchmarks, such as those isolated in the experiments of MacCrimmon (1968), the emphasis in behavioural public policy has shifted towards "common sense" policy recommendations (e.g. saving more, eating less sugar, working out more often etc.) rather than clearly articulating the normative standards they are based on.

Our historical overview ended with a discussion of agency-centric perspectives that prioritise the quality of individuals' decision-making processes over presupposing "good" behavioural outcomes. While this approach holds promise, it also raises a number of complex questions. Chief among these is the challenge of conceptualising and measuring agency as a normative standard in a way that can meaningfully inform public policy analysis and institutional reform. From an external perspective, distinguishing between genuinely acting "agentically" and merely experiencing a subjective sense of agency remains inherently difficult. A subjective feeling of agency does not necessarily equate to the objective exercise of agency. Much of the current literature on agency lacks clear normative criteria for defining what constitutes a "sufficiently good" decision-making process. Even when such criteria are proposed—such as the three conditions outlined in Section 5—it remains practically challenging to determine whether an action is preceded by adequate critical judgment, accompanied by a sufficiently large choice set, and free from manipulative third-party influences.

To avoid repeating some of the shortcomings of the nudge agenda, efforts to conceptualise and measure agency might need to move beyond the top-down perspective of the social planner, prevalent in much of the 20th century welfare economics. Instead, agency-centric approaches should explore the possibilities of "co-construction" in the form of a dialogue between citizens, behavioural economists, and public policy experts to develop normative policy benchmarks collaboratively.

REFERENCES

Alexandrova, A., and Fabian, M. (2022). Democratising measurement: or why thick concepts call for coproduction. *European Journal for Philosophy of Science, 12*(1), 7.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque : critique des postulats et axiomes de l'école américaine (The behaviour of rational man under risk: criticism of the postulates and axioms of the American school). *Econometrica 21*(4), 503–546.

Andersson, O., Lambrecht, M., & Miettinen, T. (2023). *Personal and societal conflict of distributive principles and preferences* (Helsinki GSE Discussion Paper No. 14). Helsinki Graduate School of Economics.

Banerjee, S., Grüne-Yanoff, T., John, P., and Moseley, A. (2024). It's time we put agency into Behavioural Public Policy. *Behavioural Public Policy*, 8, 789–806.

Benjamin, D. J., M. A. Fontana, and M. S Kimball (2020). Reconsidering Risk Aversion. Working Paper 28007. Working Paper Series. *National Bureau of Economic Research*.

Bell, D. E., Raiffa, H., and Tversky, A. (Eds.). (1988). *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge university Press.

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

Bernheim, B. D. (2021). In defense of behavioural welfare economics. *Journal of Economic Methodology*, *28*(4), 385-400.

Breig, Z., Feldman, P (2024). Revealing risky mistakes through revisions. *Journal of Risk and Uncertainty* 68, 227–254.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism". *University of Pennsylvania Law Review*, *151*(3), 1211-125.

Chater, N. (2022). What is the point of behavioral public policy? A contractarian approach. *Behavioural Public Policy*, 1-15.

Chater, N., and Loewenstein, G. (2023). The i-frame and the s-frame: how focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*, *46*, e147.

Colin-Jaeger, N., and Dold, M. (2025). Individual autonomy and public deliberation in Behavioral Public Policy*. Humanities & Social Sciences Communications*, 12.

Crosetto, P., and Gaudeul, A. (2024). Fast then slow: choice revisions drive a decline in the attraction effect. *Management Science*, 70(6), 3381-4165

Delmotte, C., and Dold, M. (2022). Dynamic preferences and the behavioral case against sin taxes. *Constitutional Political Economy*, *33*(1), 80-99.

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

Dold, M. (2023). Behavioural normative economics: foundations, approaches and trends. *Fiscal Studies*, *44*(2), 137-150.

Dold, M., and Lewis, P. (2023). A neglected topos in behavioural normative economics: the opportunity and process aspect of freedom. *Behavioural Public Policy*, *7*(4), 943-953.

Dold, M., van Emmerick, E., and Fabian, M. (2024). Taking psychology seriously: a self-determination theory perspective on Robert Sugden's opportunity criterion. *Journal of Economic Methodology*, 1-18.

Dold, M., and Rizzo, M. J. (2021). The limits of opportunity-only: context-dependence and agency in behavioral welfare economics. *Journal of Economic Methodology, 28*(4), 364-373.

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

Fumagalli, R. (2024). Preferences versus opportunities: on the conceptual foundations of normative welfare economics. *Economics & Philosophy*, *40*(1), 77-101.

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare 38*(4), 635–645.

Grüne-Yanoff, T., and Hertwig, R. (2016). Nudge versus boost: how coherent are policy and theory? *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science, 26*(1-2), 149–183.

Hands, D.W. (2020). Libertarian paternalism: taking Econs seriously. *Int Rev Econ* 67, 419–441 (2020).

Hands, D. W. (2024). On the (non) history of preference purification in modern economics. *Review of the History of Economic Thought and Methodology 1*(1), 1–42.

Hargreaves Heap, S. P. (2013). What is the meaning of behavioural economics? *Cambridge Journal of Economics*, *37*(5), 985-1000.

Hargreaves Heap, S. P. (2017). Behavioural public policy: the constitutional approach. *Behavioural Public Policy*, *1*(2), 252-265.

Hargreaves Heap, S. P. (2023). Mill's constitution of liberty: an alternative behavioural policy framework. *Behavioural Public Policy*, *7*(4), 933-942.

Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social Research 44*(4), 623-656.

Heukelom, F. (2014). *Behavioral Economics: A History*. Cambridge University Press.

Hertwig, R., and Grüne-Yanoff, T. (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, *12*(6), 973-986

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. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft 150*(1), 18–36.

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. and J. Snell (1990). Predicting utility. In R. M. Hogarth (Ed.), *Insights in Decision-making: A Tribute to Hillel J. Einhorn*, pp. 295–310. University of Chicago Press.

Kahneman, D. and Sugden, R. (2005). Experienced utility as a standard of policy evaluation. *Environmental and resource economics, 32*, 161–181.

Kahneman, D. and A. Tversky (1979). Prospect theory: an analysis of decision under risk. *Econometrica 47*(2), 263–291.

Kahneman, D. and A. Tversky (1984). Choices, values, and frames. *American Psychologist 39(4),* 341–350.

Kahneman, D. and A. Tversky (1996). On the reality of cognitive illusions. *Psychological Review 103*(3), 582–591.

Kahneman, D. and C. Varey (1991). Notes on the psychology of utility. In J. Elster and J. E. Roemer (Eds.), *Interpersonal Comparisons of Well-Being*, pp. 127–163. Cambridge University Press.

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. (2016). From homo economicus to homo psychologicus: the Paretian foundations of behavioural paternalism. *Œconomia. History, Methodology, Philosophy 6*(2), 175–200.

Lecouteux, G. and I. Mitrouchev (2024). The view from *Manywhere*: normative economics with context-dependent preferences. *Economics & Philosophy 40*(2), 374–396.

MacCrimmon, Kenneth R. (1968). Descriptive and Normative Implications of the Decision-Theory Postulates. In K. Borch and J. Mossin (Eds.) *Risk and Uncertainty*, 3–32. London: Palgrave Macmillan.

Marewski, J. N., and Gigerenzer, G. (2022). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, *14*(1), 77-89.

Mitrouchev, I. (2024). Normative and behavioural economics: a historical and methodological review. *The European Journal of the History of Economic Thought*, *31*(4), 533–562

Mongin, P. (2019). The Allais paradox: what it became, what it really was, what it now suggests to us. *Economics & Philosophy*, *35*(3), 423–459.

Moskowitz, H. (1974). "Effects of Problem Representation and Feedback on Rational Behavior in Allais and Morlat-Type Problems." *Decision Sciences* 5 (2): 225–42.

Nielsen, K., and J. Rehbeck (2022). When choices are mistakes. *American Economic Review* 112 (7): 2237–68.

OECD (2017). *Behavioural insights and public policy: lessons from around the world*. OECD Publishing. DOI: https://doi.org/10.1787/9789264270480-en.

Oliver, A. (2018). Nudges, shoves and budges: behavioural economic policy frameworks. *The International journal of health planning and management*, *33*(1), 272-275.

Oliver, A. (2022). Curtailing freedoms to protect freedom: regulating against behavioural-informed infringements on a fair exchange. *Journal of European Public Policy*, *29*(12), 1982-1993.

Oliver, A. (2023). *A Political Economy of Behavioural Public Policy*. Cambridge University Press.

Raz, J. (1986). *The Morality of Freedom.* Clarendon Press.

Reisch, L. A. and C. R. Sunstein (2016). Do Europeans like nudges? *Judgment and Decision-making 11*(4), 310–325.

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

Rizzo, M. J., and Whitman, G. (2019). *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*. Cambridge University Press.

Ryan, R., and DeHaan, C. (2023). The Social Conditions for Human Flourishing: Economic and Political Influences on Basic Psychological Needs.' In R. Ryan (ed.), *The Oxford Handbook of Self-Determination Theory*. Oxford: Oxford University Press.

Schubert, C. (2015). Opportunity and preference learning. *Economics & Philosophy*, *31*(2), 275–295.

Sen, A. (1999). *Development as Freedom*. Oxford University Press.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics, 69*(1): 99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138.

Slovic, P., and A. Tversky (1974). Who accepts Savage's axioms? *Behavioral Science* 19 (6): 368–373.

Spohn, W. (2011). Normativity is the key to the difference between the human and the natural sciences. In Dieks, D., (eds.). *Explanation, Prediction, and Confirmation* (pp. 241-251). Dordrecht: Springer Netherlands.

Steinberger, F. (2016). The normative status of logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition).

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

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

Sugden, R. (2019). What Should Economists Do Now? In: Wagner, R. (ed.), *James M. Buchanan: A theorist of political economy and social philosophy,* 13–37. Palgrave Macmillan

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. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1, 39-60.

Thaler, R. H. (2015). *Misbehaving: The Making of Behavioral Economics*. W. W. Norton & Company.

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

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin Books.

Tversky, A. and D. Kahneman (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology 5*(2), 207–232.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: heuristics and biases. *Science 185*(4157), 1124–1131.

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.

Viale, R. (2022). *Nudging*. MIT Press.