From Prospect Theory to Behavioural Welfare Economics

Ivan Mitrouchev

Univ Lyon, UJM Saint-Etienne, GATE UMR 5824, F-42023 Saint-Etienne, France
Université de Reims Champagne-Ardenne, REGARDS EA 6292, 51571 Reims Cedex, France

July 2020*

Abstract

Before behavioural welfare economics flourished over the last two decades, the potential normative implications of prospect theory were already a subject of discussion by Kahneman and Tversky in their seminal articles of 1979, 1981 and 1986. This article aims at clarifying some principles of behavioural welfare economics through the lens of the few normative concerns Kahneman and Tversky had in prospect theory. I show that those early references to normative analysis are informative to the methodology of behavioural welfare economics in three ways. First, they provide explanation about why the heuristics-and-biases program tends to consider deviations from rationality as *biases*. Second, they provide explanation about the practical usefulness of distinguishing descriptive and normative decision-making as two separate enterprise. Third, investigating the first stage of prospect theory helps clarifying the methodological difficulties Kahneman and Tversky had in identifying individual's underlying true preference — an important methodological problem currently faced in behavioural welfare economics. The overall argument of the paper is that contrary to the common historical transcription that the heuristics-and-biases program neglected normative concerns, I here provide a slight twist. First generation of prospect theory did not neglect normative concerns (at least not *entirely*), and the evolution of the heuristics-and-biases program during the 1990s is to be seen as a natural progression to the study of well-being measurement and policy analysis rather than a strict historical break between 'positive' and 'normative' behavioural economics.

Keywords. *behavioural welfare economics — cognitive biases — framing effect — normative analysis — prospect theory — rationality*

JEL codes. B41, D63, D90, I31

*Work in progress. I thank Niels Boissonet, Wade Hands, Cyril Hédoin, Yao T. Kpegli, Jérôme Lallement and Ramzi Mabsout for helpful comments on early versions. I also thank Jean-Sébastien Gharbi for careful reading. All mistakes remain mine.

0 Introduction

> *'By the mid-1990s, behavioral economists had two primary goals. The first was empirical: finding and documenting anomalies, both in individual and firm behavior and in market prices. The second was developing theory. ... But there was a third goal lurking in the background: could we use behavioral economics to make the world a better place? ... The time was right to take this on.'*
>
> Thaler (2015, p. 307)

Behavioural economics started as a *descriptive, explanatory and predictive* enterprise. The aim was to test whether standard decision theory — namely expected utility theory — conforms to real choices of individuals, and if not, to what extent actual choice diverges from the norms of rational choice as embodied in expected utility theory. In a series of influential contributions (Tversky and Kahneman 1973, 1974,1986, 1992; Kahneman and Tversky 1979), the heuristics-and-biases research program sought to (i) explore how heuristics lead to errors of judgement over objective probability; (ii) collect consistent and recurrent empirical findings that individuals deviate from the standard axioms of rational choice; (iii) propose a new axiomatic approach to describe/explain/predict choice from the deviations of standard decision theory. Although the normative consequences of systematic deviations from rational choice were given some attention in the early works of Kahneman and Tversky (to be discussed), the main focus of their research program was orientated to the theory of *rational choice*. That is, rather than upholding the standard separation between positive and normative economics — according to which positive economics seeks to understand and explain economic mechanisms (in other words, is about *facts*, or what *is*), and normative economics assesses policies or states of affairs (in other words, is about *ethical judgements* or what *should be*) — the meaning of 'normative' in the writings of Kahneman and Tversky referred to the standard norms of rational choice: rules for rational decision-making such as expected utility theory, logic, and Bayesian updating. There was however no particular focus on *evaluation*, *recommendation* and *prescription* of policies based on their findings.

But from the 1990s, leading behavioural economists (among them Kahneman and Thaler) have redirected a consequent part of their research to the usual meaning of 'normative economics', to be understood as the branch of economics which is about the evaluation, recommendation and prescription of policy. The main reason for this surging interest in normative economics is the accumulation of empirical evidence that individuals behave inconsistently in many ways, e.g. non-Bayesian updating (Tversky and Kahneman 1974), framing (Tversky and Kahneman 1981), self-control failure (Thaler and Shefrin 1981) and *status quo* bias (Samuelson and Zeckhauser 1988). Because of these empirical findings, some behavioural economists could not seriously take the preference-satisfaction approach of standard welfare economics anymore, according to which individuals act in a rational way so that they always know what is best for them. In fact, this latter point specifically led Kahneman and his colleagues at the beginning of the 1990s to focus on alternative measures of well-being, arguing that deriving 'true' utility from preference is questionable (Kahneman and Snell 1990; Kahneman and Varey 1991) and that paternalistic interventions may be envisaged if the State knows better what is best for individuals than individuals themselves (Kahneman 1994). There is no debate

whether behavioural economics made its own way to normative economics, a story lived and transcribed by seminal figures who drew the major lines of what constitutes most of what behavioural economics is today (Camerer and Loewenstein 2004; Kahneman 2011; Thaler 2015, 2018). However, *to what extent* behavioural economics neglected normative concerns in its early stage is a missing point in the few historical analyses of behavioural economics (Nagatsu 2015; Lecouteux 2016; Moscati 2018 [Ch. 16]).[1]

The goal of this article is to explore how far the third goal of using 'behavioral economics to make the world a better place' was actually 'lurking in the background'. Actually, Kahneman and Tversky *did* share few but notable concerns about the informative usefulness of prospect theory to normative analysis. Those include a general note about (supposed) self-acknowledged errors of reasoning by the decision maker who violates the axioms of expected utility theory (1979, p. 277), a point of discussion about individuals making errors because of framing of acts, contingencies and outcomes (1981, p. 458), and the implication of the separation between descriptive and normative decision-making for public policy (1986, p. S275). I thus aim at providing new insights to the story according to which behavioural economists had, before the mid-1990s, no concern at all in the normative implications of descriptive decision-making.[2] Although useful for having a more precise history of behavioural economics *per se*, the present historical analysis mostly serves an instrumental goal. It helps clarifying some fundamental principles in behavioural welfare economics, particularly (i) the assumption that deviations from rational choice are considered to be a prejudice against one's well-being, (ii) that descriptive decision-making is somehow informative to normative analysis, and (iii) the possibility to elicit individuals' true preferences in different contexts.

Why prospect theory? After all, it is true that behavioural welfare economics (BWE) is not tethered to exclusively one theory of the heuristics-and-biases program but instead considers the general observation that individual decision makers have cognitive biases. Two reasons explain the particular focus on prospect theory. First, one may fairly question why prospect theory (PT) — which can be considered as the descriptive decision theory that initially staged the heuristics-and-biased program — has no special status as a referent descriptive decision-making in BWE. This is concerning, knowing that several cognitive biases from which BWE is based on refer to a large extent to some components of PT — namely *reference dependence*, *loss aversion* and *probability distortion*. Second, PT is one among the most influential and popular decision theory in behavioural economics. The idea is if PT influenced in many aspects 'positive' behavioural economics and if behavioural economics switched from 'positive' to 'normative' concerns

[1] An exception is Heukelom (2014 [Ch. 4]), who provides a detailed discussion of how the descriptive/prescriptive relationship stabilised, and how the normative role of rational choice theory evolved, through the various stages of KT's research. Note that I deliberately use the fuzzy term 'normative concerns' to refer to any kind of judgements on what *should be*. As previously stated, normative analysis can either be interpreted in terms of rationality or policy (or well-being). Since the relationship between these two interpretations has never been clear (neither in the heuristics-and-biases program nor in normative economics), I leave it here for now and come back to this important point below.

[2] The sceptical reader may argue that since those normative concerns are quantitatively *few*, they may not provide strong evidence for the influence prospect theory may have had on behavioural welfare economics. To this objection, I would reply that (i) the very existence of normative concerns in first generation of prospect theory is enough to say that the heuristics-and-biases program did not *fully* neglect this aspect, and (ii) the aim of a historical reconstruction is to make those early concerns more salient, particularly when they are informative to some contemporary methodological issues (see below).

at the beginning/mid-1990s, then there is good reason to think that PT had a notable influence in the methodology of BWE.

The rest of the paper is organised as follows. Section 1 briefly introduces the components of PT by showing that they constitute an important matter of concern in BWE. Sections 2 and 3 document the early discussions KT provided about the potential normative implications of PT and compare them with the program of BWE. Those early discussions by KT are not well known. They help to understand how cognitive biases were initially considered to be normatively unacceptable (Section 2) and how the separation between descriptive and normative decision-making is useful to normative analysis (Section 3). Section 4 then provides some clarifications about the conventional assumption in BWE that framing is irrelevant to well-being by discussing the early difficulty KT had in eliciting individuals' true preferences. Section 5 concludes.

1 Prospect Theory and Behavioural Welfare Economics

BWE can be identified as the normative approach which disputes the standard view that observed preference equals to well-being due to the cognitive biases documented in the heuristics-and-biases program. It either takes the form of paternalistic interventions, which aim at improving individual well-being with almost costless impact on individual liberty (Camerer et al. 2003; Thaler and Sunstein 2003, 2009) or extending the standard welfare framework with the introduction of frames (Bernheim and Rangel 2007, 2008, 2009) or internalities (Chetty 2015; Bhargava and Loewenstein 2015). Although these works differ among themselves with respect to several features — such as which criterion of well-being should prevail (preference or choice) or whether the empirical observations of behavioural decision-making should necessarily lead to paternalistic intervention — what unifies them is the idea that cognitive biases are proper indicators of what makes individuals worse off when they make decisions. Actually, as rationality has always been the central assumption of individual behaviour in economic models many behavioural economists have considered deviations from rational choice to be 'biases', i.e. a prejudice against oneself.[3]

The key point is what behavioural economists who had a late interest in normative analysis in the 1990s originally meant by 'bias'. Did they really exclusively referred to a prejudice against oneself in terms of *rational choice* or also in terms of *well-being*? With the huge interest in 'normative' behavioural economics after the international success of *Nudge* (2009) and the establishment of Behavioural Insight Units all over the world (Halpern 2015), it appeared that the notion of 'bias' could not solely be interpreted in terms of violation of probabilistic and logical rules. Instead, there seems to be

[3]This is of course not true for all behavioural economists. In contrast with the heuristics-and-biases program (Kahneman, Slovic, and Tversky 1982), the fast-and-frugal-heuristics program holds a different normative approach by the concept of *ecological rationality* (Todd and Gigerenzer 2012 [Ch. 19]). The main point of this approach is not to consider deviations from rationality as *biases* — as Gigerenzer (1996, p. 102) puts it, 'biases are not biases'. Instead, the authors claim that some heuristics yield to 'good enough' decisions that depend on the environment in which the decision is being made. The disagreement between the heuristics-and-biases and fast-and-frugal-heuristics programs roots in fact in conflicting epistemic positions about the interpretation of probabilities (Bayesians *versus* Frequentists). See the debate between Gigerenzer (1991, 1996) and Kahneman and Tversky (1996).

a deep relationship between *rationality* and *well-being* that has never been explicit in the heuristics-and-biases program (nor in normative economics). I investigate this substantial point in Section 2. Before doing so, this first section provides a brief overview that several cognitives biases considered to be a prejudice against one's well-being in BWE actually refer to the components of PT. I then suggest two reasons that most behavioural economists interested in normative analysis have not considered PT to play a special role in the relatively new program of BWE, which (roughly speaking) emerged in the 2000s. This section is useful and necessary to start discussing the influential role PT may have had in BWE.

1.1 The Components of Prospect Theory

PT accounts for four components in decision-making: *reference dependence*, *utility curvature*, *loss aversion* and *probability distortion*. In addition, we can also consider the psychological phenomenon of *framing* (Tversky and Kahneman 1981; Tversky and Kahneman 1986) as constituting an element of the theory — although not conventionally considered as a component *per se*.

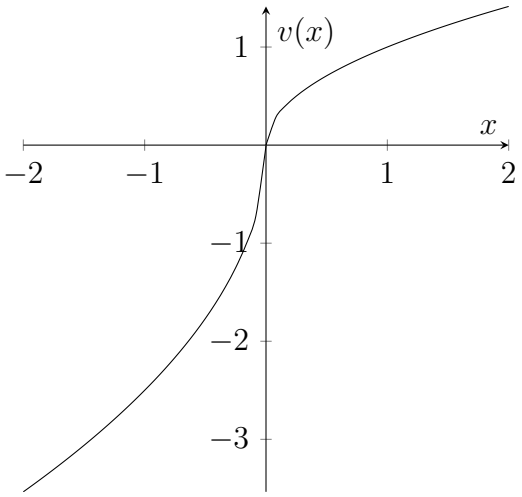
Consider first the reference point of PT. It is generally represented as the *status quo* and serves as the benchmark to distinguish gains from losses. It is assumed to be a neutral reference outcome, which is assigned the value of zero. One famous policy application which exploits this cognitive bias is the design of 401(k) saving retirement plans, where empirical evidence has shown that the 'opt-in' default option significantly increased the number of employees' enrolments (Madrian and Shea 2001; Thaler and Benartzi 2004; Bernheim, Fradkin, and Popov 2015). As Madrian and Shea (2001, p. 1181) put it, without automatic enrolment there is no reference point for the investment allocation. But with automatic enrolment, the primary reference point is unambiguously the default option. While the authors aim at improving employees' savings (and thus assuming employees will be better off by doing so), they implicitly consider that the reference point should ultimately not matter when employees make a decision between 'opting-in' and 'opting-out'. In this type of policy recommendation, if the policymaker has reasons to believe that employees would be better off by 'opting-in', the policy design should make the 'opt-in' alternative default so that employees are better off saving more than less.[4]

Consider now the curvature of the value function (utility curvature). As depicted in Fig. 1 below, the value function $v(x)$ of PT is concave above the reference point and

[4]Note that according to this normative principle, such 'bias' is exploited at the expense of individuals being unaware about their so-called 'bias'. In fact, the typical behaviour that employees stick to their initial choice can be explained by *anchoring* (Tversky and Kahneman 1974): putting heavy weight on a benchmark (e.g. the default option) in one's decision (e.g. to stick with the default option). The aim of this type of policy is not to 'de-bias' employees by making them aware about having such bias but instead to deliberately exploit their predisposition (or 'unconscious' preference) for the *status quo*. *Boosts* instead of *nudges* may be here an alternative to palliate this manipulation problem. See Grüne-Yanoff and Hertwig (2016).

convex below the reference point.[5] [6]

Fig. 1. Prospect Theory Value Function



These two first conditions refer to what Tversky and Kahneman (1992) call the principle of 'diminishing sensitivity': the more x distances from the reference point, the less impact it has on the subjective perception of the given loss/gain. For example, the perception of a loss/gain between 110$ and 120$ is less salient than the perception of a loss/gain between 10$ and 20$. Although the value function depicts choice over lotteries (or prospects) — which are most of the time binary and monetary — it can by principle also depict any choice that can be represented by gambles as long as those choices can be expressed in terms of prospects, e.g. deciding whether to eat a cake or not, to save or not or to smoke or not. This is possible because PT accounts for choice under uncertainty since second generation of PT (Tversky and Kahneman 1992). Thaler and Sunstein (2009) provide other numerous examples of non-monetary choice objects that are the matter of concern of BWE.

Loss aversion is also taken as a cognitive bias to refer to situations that are assumed to be a prejudice against oneself in BWE. According to the principle of loss aversion, losses loom larger than corresponding gains. This principle is captured by the value function $v(x)$, which is steeper for losses than for gains. For example, a loss of 1$ has more impact on the individual perception of that loss compared to a gain of 1$. Taking the same illustration of 401(k) plan designs, it is common to see loss aversion as being one

[5] The typical functional forms of the value function for gains and losses (Tversky and Kahneman 1992) are,

$$v(x) = \begin{cases} x^{\alpha} & \text{if } x \geq 0 \\ -\lambda(-x)^{\beta} & \text{if } x < 0 \end{cases}$$

where $\alpha, \beta \in [0, 1]$ are the utility curvature parameters and λ is the utility loss aversion parameter. Loss aversion holds only if $\lambda > 1$, i.e. when losses are overweighted relative to gains. For the pure sake of presentation, Fig. 1 depicts a value function $v(x)$ with $\lambda = 2.5$ and $\alpha = \beta = 0.5$ (hypothetically), where x is expressed as a deviation of the reference point $v(0)$.

[6] Note however that convexity in the loss domain is not required in the axiomatisation of prospect theory. It used to be a prediction made by Kahneman and Tversky (1979) and Tversky and Kahneman (1992) but some empirical studies suggest the possibility of concavity in the loss domain (Abdellaoui, Bleichrodt, and L'Haridon 2008).

source (among others) of the *status quo* bias. Consider the choice between 'opting-in' and 'opting-out' as a mixed gamble where the outcome is uncertain. That is, one does not know with certainty the utility which will be gained by her future being from the decision taken by her present being. As a consequence, employees typically prefer to stick to their initial choice.[7] Another example of loss aversion can be provided in the market experimental designs of Kahneman, Knetsch, and Thaler (1990), where subjects were given two different goods and asked how much they were willing to sell/buy their good in exchange of the other good. Due to *endowment effect* (Thaler 1980) — the observation that individuals often demand much more to give up an object than they would be willing to pay to acquire it — the common conclusion is that they experience loss aversion. In those experiments, such effect was measured by the discrepancy between willingness to accept (WTA) and willingness to pay (WTP) for the other good. When the WTA was significatively greater than the WTP, the authors inferred that individuals experienced strong loss aversion. Such behaviour is assumed to be 'irrational' by the authors because they observed the same behaviour for individuals who were given the other good and because through those experiments, they empirically falsified the standard argument that the market environment eventually makes such irrational behaviour disappear by learning opportunities.[8]

Probability weighting also strongly refers to some cognitive biases documented in BWE. In PT, probabilities are replaced with decision weights that involve probability weighting function $w(p)$. The probability weighting function is an increasing function of p, but not a probability. It represents the psychological perception of a probability, i.e. the psychological weight individuals put to the realisation of events. One important property of the function w is 'subcertainty', which means that low probabilities are overweighted, moderate and high probabilities are underweighted, and the latter effect is more pronounced than the former (see Fig. 2 below).[9]

[7] Such phenomenon is more generally explained by Schwartz (2004 [2016, Ch. 6]) in terms of opportunity cost. When evaluating two options that seem both attractive (or to which one is relatively indifferent), individuals fear of making the 'wrong' choice because of loss aversion (i.e. missing the opportunity of making the other choice) and therefore prefer either to choose nothing or to stick to the default option.

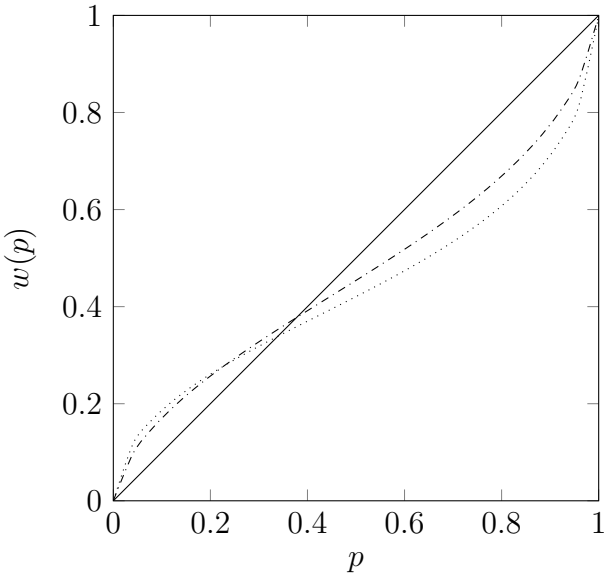
[8] Harrison and Ross (2017) provide an interesting criticism to the assumption that loss aversion is a prejudice against oneself. In their words, 'λ is a response operator, most naturally interpreted as reflecting a sentimental influence on behavior and cognition. To the extent that a person experiences direct sentimental disutility from losses *per se*, whenever she interprets an outcome as a loss, then it seems straightforwardly presumptuous to maintain that a policy-maker should override this aspect of her psychology' (p. 8). Confusingly, the interpretation of loss aversion in terms of displeasure may appear contradictory to the principle of constructing the value function, which does not represent *hedonic states* but *choices* over prospects. I come back to this point below.

[9] One typical functional form of the weighting function (Tversky and Kahneman 1992) is,

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}}$$

where $\gamma \in [0, 1]$ represents the distortion of probability parameter. The parameter γ is typically different between positive (w^+) and negative (w^-) prospects. For the pure sake of presentation, Fig. 2 depicts a decision weighting function for estimated parameters by Tversky and Kahneman (1992, p. 312) $\gamma = 0.69$ for negative prospects (thick dotted line) and $\gamma = 0.61$ for positive prospects (thin dotted line).

Fig. 2. Prospect Theory Decision Weighting Function



Examples of probability distortions that may affect individual well-being are the purchase of insurances after a flood (Thaler and Sunstein 2009), extended warranties and state lotteries (Camerer et al. 2003). As captured by PT, individuals typically overweight small probabilities for gains — e.g. state lottery gamblers are typically optimistic about their chance to win — and for losses — e.g. insurance-buyers typically think their house to have more chance to flood after experiencing this tragic event, and warranty-buyers typically think their purchased good to have more chance to break than what the actual odds are. This attitude is generally and jointly captured by the two functions $v(x)$ and $w(p)$, which suggest risk aversion for gains and risk seeking for losses of high probability, and risk seeking for gains and risk aversion for losses of low probability.[10]

Lastly, framing is extensively discussed in KT (1986) as a cognitive bias that should not affect individual choice. A well-known example is the cafeteria-director introductory problem of Thaler and Sunstein (2003, 2009), whose aim is to displace apples and cakes on the counter in a way that the consumption of apples increases, this without restricting the liberty of those who would like to consume a cake (e.g. displacing the apples slightly in front of the cakes). As framing is the subject of discussion in the last section, I leave it here for now and simply end up with my comment that it is, strictly speaking, an important matter of concern in BWE, which deals with what Thaler and Sunstein (2009) would call a 'choice architecture': the indirect way of influencing individual choice through the framing of that choice (e.g. displacing the apples slightly in front of the cakes).

My point is the following. Knowing that almost every component of PT are directly the matter of concern of BWE, it seems surprising that almost no contribution in BWE

[10]As Hands (2020) puts it on Camerer et al.'s (2003) example of warranties, one problem of associating probability distortion with a prejudice against oneself is that 'it may be that people make mistakes when they buy such warranties and they do not realize how unlikely such expenses are, but it may be that even fully informed they would still do it (i.e., it is not a mistake for them), they just put a high value on peace of mind' (sec. 2). This suggests the question of whether the examples provided by Thaler and Sunstein (2009) and Camerer et al. (2003) only deal with situations where individuals would *not know* the probabilities associated to given events. But this is not what the authors seem to mean.

mention PT as *the* descriptive model of decision-making from which evaluation, recommendation and prescription of policy can be derived from. Strictly speaking, only Bleichrodt, Pinto, and Wakker (2001) and Pinto-Prades and Abellan-Perpiñan (2012) aim at deriving normative assessments from PT using expected utility theory as the correct model of normative decision-making — or as the first class of authors name their paper, 'making descriptive use of prospect theory to improve the prescriptive use of expected utility'. But for the rest of BWE, the reference to PT has either been completely neglected or at best briefly referred to in footnotes (see Camerer et al. 2003, pp. 1215-1216). I shall posit two reasons that may explain the negligence of prospect theory as the adequate positive model for normative analysis.

1.2 Two Reasons That Prospect Theory Holds *a Priori* No Special Role in Normative Analysis

First, there seem to be no particular reasons for BWE to be derived from only *one* model of descriptive decision-making because all kinds of studies that would supplement our knowledge about how individual decision makers actually choose could in some way inform normative analysis, from psychology to neuroscience. This view is for example explicitly the one adopted by Camerer, Loewenstein, and Prelec (2005) and Camerer (2008). According to these authors, not only enhancing our knowledge from other sciences about how individuals choose may inform normative analysis, but before all positive analysis.[11]

Second, the value function of PT depicted above represents *decision* utility (based on choice over lotteries) but not *experienced* utility (the hedonic state of experiencing something). Consequently, at the beginning of the 1990s Kahneman and his colleagues had to find alternative measures of well-being that are not related to what individuals choose but instead to what they experience (Kahneman and Snell 1990; Kahneman and Varey 1991). Since the value function represents individuals' *choice* and since behavioural welfare economists assume that choice (or observed preference) is not equal to well-being, it follows that the value function cannot measure individual well-being. At best, the value function can inform the policymaker on how individuals actually choose, so that his policy recommendation may be justified from this observation. What is however interesting to note is that even though hedonic psychology played no role in the development of PT, KT interpreted loss aversion in terms of a pleasure/displeasure metric:

> 'The displeasure associated with losing a sum of money is generally greater than the pleasure associated with winning the same amount' (Tversky and Kahneman 1981, p. 454)

It would be perhaps unwarranted to say that KT had in mind in the value function an intensity of pain and pleasure, similarly to what Kahneman, Wakker, and Sarin (1997) lately meant by a hedonic measurement of well-being. Again, the value function depicted in Fig. 1 represents the *decision* utility associated with possible outcomes of the decision at hand, not the *experienced* utility of the reference situation. For the purpose of PT, it was useless to assume that the value function had something to do with a pleasure metric. But there was no reason to consider the opposite either. From the point

[11] See the important epistemic debate about how preferences should be represented in economics (either 'behavioural' or 'mental') and the ethical debate about whether economics should have welfare implications on policymaking in Caplin and Schotter (2008).

of view of behavioural welfare economists, the analogy between the value function and the pleasure/displeasure of gains/losses may have intuitive appeal for practical purpose, typically to justify that loss aversion is a prejudice against one's well-being. But then this interpretation of loss aversion in terms of displeasure seems quite contradictory with the principle that decision utility is different from experienced utility, a distinction already well recognised by KT (1981, p. 458). In fact, Kahneman (1999, p. 18) particularly discussed the possibility of loss aversion in experienced utility, so that the value function depicted above may have a similar shape than the experienced utility function (which functional form was at the time empirically unknown). The similarity of behaviour between decision utility and experienced utility was lately subject to an empirical test proposed by Carter and McBride (2013), who actually found a similar S-shape for both functions. Their result ultimately suggests that although both concepts of decision utility and experienced utility make sense from a theoretical viewpoint, the two are related at a fundamental level.

Having set out that the components of PT are intimately related to some cognitive biases that are of considerable matter of interest in BWE, the next section discusses the influence PT may have had on some known principles in BWE. I discuss the assumption of true preference and the practical usefulness for normative analysis of strictly separating descriptive from normative decision-making.

2 Cognitive Biases as Normatively Unacceptable

KT (1979) initially proposed PT as an alternative descriptive model of decision-making under risk to expected utility theory. In their series of seminal articles, the normative interpretation of rational choice for descriptive purpose was a central point of criticism towards standard decision theory. KT (1986) particularly argued that the violation of two essential axioms of rational choice — dominance and invariance — cannot provide a satisfactory normative representation of descriptive decision-making under risk.[12] The main criticism KT addressed to expected utility theory was specifically being a normative model of decision-making — what decision makers *should* do — instead of being a positive model of decision-making — what decision makers *actually* do. The argument that PT departs from the normative interpretation of rationality in standard decision theory was then seemingly promoted in their proposition of cumulative PT (1992, pp. 297, 301, 317). Evidently, KT presented PT as a model of decision-making that was free from any normative concern. By 'normative', KT meant how the concept is commonly deployed in decision theory — that is to say, the way decision makers would like to choose, typically by the norms of rational choice defined under a set of axioms. But although such departure from the normative concern of descriptive decision-making was explicit in their proposition of PT, they also gave few notes of discussion in their articles of 1979, 1981 and 1986 about the potential implications of PT for the evaluation of states of affairs and recommendation/prescription of policies.

[12] *Dominance* states that if one option is better than another in one state and at least as good in all other states, the dominant option should be chosen. *Invariance* states that different representations of the same choice problem should yield the same preference.

2.1 Anomalies and True Preferences

In their original proposition of PT, KT first suggested a normative principle familiar with BWE when discussing the potential normative implications of subjects who would deviate from 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, p. 277)

The term 'anomalies' is here used to characterise choices that are not consistent with expected utility theory but which are taken into account by PT.[13] With the growing interest of behavioural economics towards normative analysis in the 2000s, 'anomalies' took however another twist. The term refers not only to deviations from the axioms of expected utility theory but also to 'unacceptable' choices that individuals would have corrected had they been initially in full possession of their computational skills, their willpower, and had they been well informed (or been provided an *ex-post* feedback) (Thaler and Sunstein 2003, p. 175). Typically, shall the decision maker be informed about her (supposedly) erroneous choice — i.e. the one that is not optimal according to her interests — it is assumed that she would correct her choice by choosing according to her (supposedly existing) preferences that are undistorted from cognitive biases. Those 'unbiased' preferences are often called *true* preferences. Before the existence of true preferences became a common assumption in BWE, KT wisely questioned the status of observed preferences as a normative criterion when those preferences are judged to be 'incoherent'.

> 'The present work has been concerned primarily with the descriptive question of how decisions are made, but the psychology of choice is also relevant to the normative question of how decisions ought to be made. In order to avoid the difficult problem of justifying values, the modern theory of rational choice has adopted the coherence of specific preferences as the sole criterion of rationality. This approach enjoins the decision-maker to resolve inconsistencies but offers no guidance on how to do so. It implicitly assumes that the decision-maker who carefully answers the question "What do I really want?" will eventually achieve coherent preferences. However, the susceptibility of preferences to variations of framing raises doubt about the feasibility and adequacy of the coherence criterion.' (Tversky and Kahneman 1981, p. 458)

With the observation that individuals violate several axioms of rational choice, the central question is what normative status should coherent preferences have in BWE. In microeconomic theory, a coherent preference is a preference that satisfies several conditions of rational choice, principally completeness, transitivity, context-independency and stability over time. In addition to those conditions, and in KT's terms, a coherent preference includes the non-violation of axioms of expected utility theory such as dominance, invariance and the important independence axiom (Allais 1953). In standard

[13] Lately, the term was associated to a series of articles in the *Journal of Economic Perspectives* to report new empirical observations which violate the standard rational choice paradigm. In the words of Thaler (1987) who holds the first number of the series, 'an empirical result is anomalous if it is difficult to "rationalize," or if implausible assumptions are necessary to explain it within the paradigm ... that agents have stable, well-defined preferences and make rational choices consistent with those preferences in markets that (eventually) clear' (p. 197).

welfare economics, the welfare principle of preference-satisfaction is to take exclusively *coherent* preferences to be normatively relevant. This principle is implicitly grounded on two features. The first is the ethical theory of welfare economics, according to which preference-satisfaction decently indicates what makes individual better off. The theory states that if an individual prefers x to y, this preference makes it the case that x is better for her than y (Broome 2009, pp. 10-11).[14] The second is that in order to identify which preferences count as welfare-relevant and which do not, standard economics does not have a concept of well-being that fits adequately with this account but *rationality*.

2.2 Rationality Rescued

The relationship between rationality and well-being has however always been fuzzy, not only in the heuristics-and-biases program but also in normative economics. One reason is that rationality is a term that can contains several meanings, e.g. 'instrumental', 'procedural', 'bounded' or 'substantive', and the interpretation of it solely depends on the economist's subject of interest. Another reason is albeit rationality can take many meanings, its western interpretation of being 'the right way to think' is appealing to the scientist so he does not think he has to deal with the difficult problem of justifying ethical values — an enquiry that standard welfare economics is reluctant to. This position about welfare economics is for example taken by Bernheim (2016), who states that assessing whether certain moral judgements are flawed is not the task of the conventional economic framework, which instead 'seeks to assess well-being without factoring in these types of moral considerations' (p. 18). Many economists have however argued that value judgements are simply inevitable in normative analysis and that standard concepts such as Pareto efficiency — the central criterion of welfare economics that is often considered to be weakly value-loaded — is not necessarily weak when compared with different ethical views. In fact, most behavioural welfare economists defend their approach with the argument that normative economics does not involve ethical/moral judgements but the 'right way to think', which is governed by the laws of logic (and not by the laws of ethics — whatever that might mean). Thaler's (2015) introduction of PT in his popular *Misbehaving* makes this view explicit.

> 'The organizing principle was the existence of two different kinds of theories: normative and descriptive. Normative theories tell you the right way to think about some problem. By "right" I do not mean right in some *moral* sense; instead, I mean *logically consistent*, as prescribed by the optimization model at the heart of economic reasoning, sometimes called rational choice theory. That is the only way I will use the word "normative" in this book.' (Thaler 2015, p. 25 — my emphasis)[15]

Now if the reader seriously takes this meaning of 'normative', how is she supposed to understand the use of descriptive decision-making for making 'the world a better place?'. It appears that behavioural welfare economics can hardly live a double life: on the one hand, recommending public policies such as saving more (a behaviour that has obviously nothing to do with the rules of logic) and on the other hand, correcting individuals'

[14] As Broome (1991, p. 4) puts it, this is considered to be true even if nothing in the definition of utility — the value of a function that represents an individual's preferences — suggests that a preferred alternative is necessary better for the individual.

[15] See also Thaler (2018): 'By normative here I mean a theory of what is considered to be rational choice (rather than a statement about morality)' (p. 1267).

rational errors labelled as 'bias' only in accordance with the rational benchmark.[16] To put it simply, it seems straightforward that the 'Human *versus* Econ' distinction in Thaler and Sunstein (2009) is strikingly used as an analogy between how individuals do and ought to behave, however both in terms of logical consistency *and* in terms of what is best for them. Yet any statement which provides an answer to the old Socratic question of what one *should* do is inevitably grounded (for the worse or the best) in the field philosophers call ethics. We may then see the following succession in the evolution of the heuristics-and-biases program.

1. *'Early' heuristics-and-biases program (before 1990s)*. Cognitive bias = psychological state considered to be a prejudice against individuals in terms of *rational choice*, without saying anything on what is 'good' or 'bad' for them.

2. *'Mid' heuristics-and-biases program (after 1990s)*. Cognitive bias = psychological state considered to be a prejudice against individuals in terms of *well-being*, i.e. what is good or bad for them in the ethical sense.

3. *'New' heuristics-and-biases program (since 2000s)*. Psychological state considered to be a prejudice against individuals in terms of rational choice = psychological state considered to be a prejudice against individuals in terms of well-being. That is, rationality and well-being are now conflated.

Thaler (2015) gives the reader a hint in what is never explicit in Thaler and Sunstein (2009), but which makes perfect sense with the author's presentation of PT as a *positive* (by contrast to a *normative*) decision theory:

> 'With prospect theory, Kahneman and Tversky set out to offer an alternative to expected utility theory that had no pretense of being a useful guide to rational choice; instead, it would be a good prediction of the actual choices real people make. *It is a theory about the behavior of Humans*.' (Thaler 2015, p. 29 — my emphasis)

The author could have safely continued this line with '... while expected utility theory is a theory about the behavior of Econs'. Some economists do explicitly recognise the separation of PT and expected utility theory as the 'actual' and 'right' models of decision-making (Bleichrodt, Pinto, and Wakker 2001; Pinto-Prades and Abellan-Perpiñan 2012). The authors make it very clear that by assuming PT as the descriptive model of decision-making and expected utility as the normative model of decision-making, PT is the *actual* way of behaving and expected utility is the *right way* of behaving, both logically and morally. This important point has nonetheless always been ambiguous in the heuristics-and-biases program.[17]

[16] In addition to Thaler (2015, 2018), this second life is very often the one privileged by behavioural welfare economists. For example, Dhami (2016) specifically introduces BWE by the following awareness: 'the terms biases and misperceptions only make sense, relative to the rational benchmark in neoclassical economics. There should be no presumption that, in any absolute sense, the actual behavior of humans should either be termed as a bias or a misperception' (p. 1577).

[17] The approach of Pinto-Prades and Abellan-Perpiñan (2012) is actually proposed to palliate the fuzzy 'benchmarks' of libertarian paternalism, according to which individuals are better off if they had complete information, unlimited cognitive abilities and no lack of willpower. In the authors' account, by setting loss aversion and probability distortion as prejudices against one's well-being, their approach has the merit of clearly measuring the discrepancy between individuals' actual choice and how they ought to choose.

In short, one may not necessarily be free from justifying value judgements as KT (1981) suggested, even if one preserves the concept of coherent preference as a normative criterion. Interestingly, by assuming that the satisfaction of coherent preferences is what makes individuals better off, this early view of KT is in fact well in line with standard welfare economics. The only difference with standard welfare economics is that coherent preferences are now disentangled from observed preferences (the ones that are subject to cognitive biases, e.g. framing). From a welfarist perspective, it can be said that BWE takes true preferences — which appear to be well aligned with coherent preferences — as the informational basis of the welfare-relevant domain. Before this principle became largely popular in BWE, it was well summarised by KT in the following four points when individual decision makers are subject to framing.

> 'Individuals who face a decision problem and have a definite preference (i) might have a different preference in a different framing of the same problem, (ii) are normally unaware of alternative frames and of their potential effects on the relative attractiveness of options, (iii) would wish their preferences to be independent of frames, but (iv) are often uncertain how to resolve detected inconsistencies.' (Tversky and Kahneman 1981, p. 458)

Note how (iii) echoes with the concept of true preference and (ii) and (iv) provide the social planner a legitimate status in behavioural paternalism when individuals are unable to 'purify' or 'optimise' their preferences themselves.[18]

3 The Separation between Descriptive and Normative Decision-Making

A second insight about the influence PT may have had on BWE regards KT's discussion on the separation between descriptive and normative decision-making, where the former is informative to the latter. The main argument of KT (1986) is since the essential axioms of dominance and invariance are violated by empirical evidence, normative decision theory cannot provide an adequate descriptive model of decision-making under risk. But the separation between normative and descriptive model of decision-making does not neglect anything from the informative usefulness of PT to normative analysis. Quite the contrary. Documenting several psychological biases that individuals can experience may constitute a source of information for the social planner about which choices could be considered to be misleading according to their own interests.

> 'the normative and the descriptive analyses of choice should be viewed as separate enterprises. ... To retain the rational model in its customary descriptive role, the relevant bolstering assumptions [that substantial violations of the standard model are (i) restricted to insignificant choice problems, (ii) quickly eliminated by learning, or (iii) irrelevant to economics because of the corrective function of market forces] must be validated. Where these assumptions fail, it is instructive to trace the implications of the descriptive analysis (e.g.,

[18]Note also that the term 'preference purification' (Hausman 2012) of the seminal inner rational agent critique of Infante, Lecouteux and Sugden (2016a, 2016b) may not be appropriated. In terms of microeconomic theory, it would actually be more accurate to say that individuals, whose aim is to optimise their utility function subject to constraints, fail to *optimise* by making cognitive errors. For example, due to framing they ultimately deviate from their demand functions. But 'pure' preferences is not a terminology endorsed by economists to characterise coherent (or rational) preferences. That is, nothing seems 'impure', strictly speaking, to deviate from coherent preferences. See Hands (2020) for an assessment of libertarian paternalism by 'taking Econs seriously'.

> the effects of loss aversion, pseudocertainty, or the money illusion) for public policy, strategic decision-making, and macroeconomic phenomena.' (Tversky and Kahneman 1986, p. S275)

Two arguments may support the view that the value function of PT can be somehow informative to normative analysis.

3.1 The Empirical Adequacy of Prospect Theory

First, by improving the empirical validity of a descriptive model of decision-making (here PT), social planners or policymakers can have relevant information about what may constitute an 'erroneous' choice. As Bernheim and Rangel (2007, 2009) and Bernheim (2016) state, this can be done by identifying the operational misunderstanding of the relationship between means and outcomes. According to the authors, a psychological process could unambiguously be labelled as a 'mistake' if it refers to objective properties of human cognitive abilities, typically the observation, attention, memory, forecasting and learning processes of individuals. Note however that although focusing exclusively on 'objective' properties of human cognitive abilities may appear weakly value-loaded, it still inevitably involves Bernheim and Rangel to make a value judgement about what a 'good' and a 'bad' choice is. In their account, a 'good' choice is a choice made with full cognitive capacities, just as in all other approaches in BWE.

But there is perhaps a fundamental point that may actually not play out in favour of using PT as the right model of decision-making for normative analysis. If it appears that PT is in fact *empirically inadequate* then all the rhetoric under which individuals are 'biased' when they make decisions in conflict with rational choice theory falls apart. This point is specifically the subject matter of Harrison and Ross (2017). The authors provide important criticisms of the empirical adequacy of PT by discussing the estimation of all parameters $\lambda, \alpha, \beta, w^+, w^-$, the test of the theory on hypothetical choices, the violation of asset integration and econometric methods that accommodate for individual heterogeneity. For example, PT considers a function $V(v_1, w_1; \ldots; v_n, w_n)$ such that values are assigned to gains and losses rather than final assets in response to the violation of the *asset integration* axiom, which states that a prospect $(x_1, p_1; ...; x_n, p_n)$ is acceptable at asset position w if and only if $U(w + x_1, p_1; ...; w + x_n, p_n) > u(w)$.

As the authors put it, violation of asset integration is however not always observed in empirical tests of decision theories (pp. 6-7). In an experimental design where cumulative PT was tested against expected utility theory and rank dependent utility theory, Harrison and Swarthout (2016) found that subjects *do* asset integrate. The study of Harrison and Ross (2017) results in a sceptical evaluation of (cumulative) PT as the correct descriptive theory for BWE. They argue that rank dependent utility theory (Quiggin 1982) is instead more appropriate as a descriptive model of decision-making, especially when expected utility theory is to be considered as the relevant normative standard. The main argument of Harrison and Ross (2017) is if BWE relies on the assumption that PT is a good descriptive model of decision-making to make normative assessments (such as in the approaches of Bleichrodt, Pinto, and Wakker (2001) and Pinto-Prades and Abellan-Perpiñan (2012)), it is sufficient to show that PT is not an empirical adequate model of decision-making in order to make the program of BWE fail.

3.2 The Non-Delimitation of Choice Objects

Second, since PT in particular and models of decision-making in general are not delimitated to a specific range of objects, it allows to use a model of descriptive decision-making for any kind of normative assessment involving gains and losses. As initially stated by KT, PT is a powerful model of decision-making which applies not only to monetary gains but extends to other objects of choice, including ethical choices about number of lives that could be lost/saved related to a policy decision.

> 'Although the present paper has been concerned mainly with monetary outcomes, the theory is readily applicable to choices involving other attributes, e.g., quality of life or the number of lives that could be lost or saved as a consequence of a policy decision.' (Kahneman and Tversky 1979, p. 288)

When originally developed, the empirical evidence of PT was based on few experiments on choices regarding monetary outcomes, the gain of travel trips and the loss of human lives (Kahneman and Tversky 1979; Tversky and Kahneman 1981). It then broadened to a wide domain of applications including tax lottery cases (Chang, Nichols, and Schultz 1987), intertemporal choice (Loewenstein 1988), portfolio investment (Benartzi and Thaler 1995), health (Attema, Brouwer, and L'Haridon 2013; Attema, Bleichrodt, and L'Haridon 2018) and investment on stock market (Barberis, Mukherjee, and Wang 2016).[19] The idea is if descriptive decision-making can apply to other non-monetary decisions and also to ethical choices (e.g. the number of lives saved), there is no intuitive objection for not considering PT as a source of information for normative analysis, which the latter is largely concerned with non-monetary outcomes.

4 Framing as Irrelevant to Well-Being

This fourth and last section provides a last piece of the puzzle in my investigation of the influence PT may have had on BWE. Perhaps one of the most important assumption made in BWE is that frames are irrelevant to well-being (Bernheim and Rangel 2007, 2008, 2009). To provide an illustration, consider an individual who would like to commit to a diet but when faced to the choice set {*cake*, *apple*}, chooses the cake over the apple. According to standard decision-making, the act of choosing the cake reveals a preference for the cake. But as noted previously, BWE considers observed preferences not to be necessarily equal with well-being and therefore with one's true preferences. An individual may reveal a preference for the cake over the apple while being truly better off with the apple (the healthier option) over the cake (the less healthy option). Due to the way both options are presented, there is a chance that the individual chooses the cake. But had she not been distorted by cognitive biases (here framing), she would have chosen the apple. Framing is a vocabulary originally used by KT (1984, p. 343) that actually designates the *isolation effect*: how individuals choose depending on how the choice problem is framed (violation of invariance).

The invariance axiom states that the preference order between prospects should not depend on the manner in which they are described. In particular, two versions of a choice

[19] For a review of the studies which provide empirical support to PT from 1979 to 1995, see Edwards (1996, pp. 22-32). For a more recent review, see Barberis (2013).

problem that are in fact equivalent when shown together should elicit the same preference, even when they are shown separately. As an illustration of violation of invariance, consider the following choice problem proposed by KT (1981, p. 454).[20]

Problem 1 [N = 150]
A: a sure gain of 240$ [84%]
B: 0.25 chance to gain 1000$ and 0.75 chance to gain nothing [16%]

Problem 2 [N = 150]
C: a sure loss of 750$ [13%]
D: 0.75 chance to lose 1000$ and 0.25 to lose nothing [87%]

In the two problems presented above, it is specifically because 'in many situations ... the decision maker does not have the opportunity to discover that his preferences could violate decision rules that he wishes to obey' (1979, p. 277) that KT suggested to use 'neutral' frames (whenever possible) in order to characterise a benchmark for identifying those mistakes. The method KT initially proposed in order to know in which frame individuals make a mistake/error is to set a third frame in addition to the previous two where both of the prospects are combined so that individuals have a transparent version of the choice problem (the neutral frame):

Problem 3 [N = 86]
A & D: 0.25 to win 240$ and 0.75 to lose 760$ [0%]
B & C: 0.25 to win 250$ and 0.75 to lose 750$ [100%]

When the prospects were combined and the dominance of the second option became obvious, all the respondents chose the superior option. In BWE, the policymaker/social planner would then only take the expressed preferences in the third choice frame as normatively relevant. This nonetheless requires to establish a rule (or criterion) that tells us why the separate choices of A & D are better than the separate choices of B & C. The implicit argument put forward by KT is that *Problem 3* provides unambiguous normative relevance because respondents unanimously preferred B & C [100%] to A & D [0%] (*unanimity*). Had at least one respondent preferred A & D to B & C, another possible normative rule would be *majority*: the option that should provide normative guidance to the policymaker is the one chosen by most individuals. Another possible normative rule is what should be preferred independently of individuals' responses, e.g. more money to less (*dominance*). That is to say, even if most respondents preferred A & D to B & C, this rule would state that it would still be a mistake to have done so because A & D provides less gains that B & C.

Now what if a third 'neutral' frame can simply not be proposed? Consider for example the following choice problem in two different frames (KT 1981, p. 453), where subjects were asked to choose a treatment against an Asian disease which is expected to kill 600 people (*Problem 4* and *Problem 5*).

[20]The number of subjects for each frame and the frequency of responses are in brackets. The following choice problem relates to framing of *acts*, but invariance is also violated in framing of *contingencies* and *outcomes*. As BWE is mostly concerned with mistaken/erroneous *choices*, the framing of acts is the relevant framing effect to be discussed here.

Problem 4 [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%]

Problem 5 [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%]

Contrary to *Problem 3*, there is no third 'neutral' frame on which an external viewpoint can rely on in order to elicit individuals' true preferences. This is because no combination of alternatives can be presented in a third frame so that one transparently dominates the other. Indeed, since *Problem 4* and *Problem 5* yield identical outcomes, the combination of A & D and B & C are also identical. Crucially, it remains an open question which of the two frames/contexts (either *Problem 4* or *Problem 5*) should be here normatively relevant. KT were actually perfectly aware that in some situations, a third alternative which would provide normative guidance can hardly be known.

> 'In some cases (such as problems [1, 2 and 3]) the advantage of one frame becomes evident once the competing frames are compared, but in other cases (problems [4, 5]) it is not obvious which preferences should be abandoned.' (Tversky and Kahneman 1981, p. 458)[21]

This point is also well acknowledged in the literature:

> 'Preference reversal raises a very awkward question: if choices and valuations reveal different preference orderings, which, if either, reflect true preference? Without an answer to this question we do not know on which elicitation methods, if any, we can rely for obtaining sound preference data.' (Braga and Starmer 2005, p. 60)

> 'it is well-established that the preferences that are revealed in people's choices over pairs of options differ systematically from those that are revealed in their separate monetary valuations of the same options, but it is far from clear which (if either) of these preferences is "correct".' (Sugden 2010, p. 54)

Some attempts to identify mistakes have been proposed in BWE, but those do not provide a clear answer to the issue of knowing which frame is normatively relevant. One attempt made by Bernheim (2016) is to characterise a mistake by two formal properties. The first is that an individual may pick the wrong alternative because she was not fully informed — what Bernheim (2009) calls the 'characterisation failure'. The second condition is that had the individual been fully informed, there is one alternative over the others that she would choose with certainty. It is however implicit in Bernheim's characterisation of a mistake that the social planner is able to identify the neutral frame in which the individual would choose her preferred alternative with certainty. But again, on which meta-criterion such neutral frame is supposed to be identified (unanimity, majority, dominance etc.) and what to do if a choice problem such as the one presented in *Problem 4* and *Problem 5* is not apt for providing a dominant option, remain open questions. Some may argue that no elicitation method can be satisfactory unless we have individuals' own explicit acknowledgement about making an error. For example, after presenting individuals *Problems 1, 2* and *3*, we could ask them whether they actually

[21] The numeration of the choice problems are changed in this quote in order to fit the ones used in the present paper.

think they have made an error by choosing A & D. Of course, no behavioural welfare economists would be opposed to this principle, but the issue is specifically to have a normative rule when those *ex-post* feedback are unavailable (see Bleichrodt, Pinto, and Wakker (2001)). Once again, it is striking to see that this important problem of BWE was already well recognised in KT's early works.

5 Conclusion

This article shows that although playing a marginal role in the development of the theory, the early normative concerns KT had about PT appeared to had a significant influence in the methodology of BWE. I have developed three main points which support this view.

First, PT may provide intuitive information about what could be considered as an 'acceptable' or 'unacceptable' choice when a third party observes subjects who violate several axioms of rational choice such as dominance and invariance (KT 1986), or when they observe psychological phenomena they label as 'biases' (by reference to rational choice theory) such as *status quo* or loss aversion. This relies on the important assumption of true preference and by making value judgements about what is considered to be a 'good' or a 'bad' choice. It requires to continue with the old tradition of standard welfare economics, which considers a coherent preference to be normatively relevant — and a true preference seems to be nothing more than a coherent preference, i.e. a preference that satisfies several conditions of rational choice. The aim of BWE is then to assess policies based on individuals' true preferences.

Second, a decision theory is specifically powerful because it does not specify the object of choice it is concerned with, and PT is far from being an exception. The idea is (i) if descriptive decision-making can provide information about how individuals make decisions and (ii) if it is applicable to any type of choice, it 'intuitively' follows from (i) and (ii) that PT could be informative towards any kind of choice that affects one's well-being.[22] In other terms, it seems that we could build rules on how individuals *ought to* choose based on behavioural observations, e.g. violation of dominance and invariance. But on which ethical premise we should judge a choice to be 'good' or 'bad', or whether prospect theory is actually empirically adequate, are up to question.

Third, the conventional assumption that framing is irrelevant to well-being can be understood with the violation of the invariance axiom that is implicitly taken as a decent normative benchmark in BWE. For practical purpose, it may be convenient for BWE to keep the assumption of context-independency as normatively relevant. This however comes up with all the methodological difficulties associated with the assumption that frames are irrelevant to well-being, e.g. the criteria to judge what counts as a welfare-relevant frame (majority, unanimity, etc.), or even the 'no-frame' problem (that no choice situation is context-independent) — a problem that has not been discussed here.

We can then draw few lessons from the influence PT may have had on BWE. First,

[22] I say 'intuitively' because nothing says that such *is-ought* relationship is a logical implication. This would require to discuss how Hume's is-ought problem can be somehow bypassed — which is an enquiry that is outside the scope of the present paper.

contrary to the historical transcription that the heuristics-and-biases program neglected normative concerns, the added value of my analysis is that this historical transcription is not entirely true as the quotes in KT (1979, 1981, 1986) show. Second, with the switching from positive to normative concerns of the heuristics-and-biases program in the 1990s, BWE is more likely to be seen as a natural extension of the heuristics-and-biases program rather than a new area of research *per se*. This is because the interpretation of a prejudice against oneself was never entirely clear, i.e. either referring to rationality or well-being. Third, some methodological issues of BWE — namely the elicitation of true preference — become more salient when we confront them with the early methodological difficulties KT had in making such enterprise possible at all. Surprisingly, these methodological issues were already striking even though the authors had at the time no significant interest in normative analysis.

References

Abdellaoui, M., H. Bleichrodt, and O. L'Haridon (2008). A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty 36*(3), 245–266.

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.

Attema, A. E., H. Bleichrodt, and O. L'Haridon (2018). Ambiguity preferences for health. *Health Economics 27*(11), 1699–1716.

Attema, A. E., W. B. F. Brouwer, and O. L'Haridon (2013). Prospect theory in the health domain: a quantitative assessment. *Journal of Health Economics 32*(6), 1057–1065.

Barberis, N., A. Mukherjee, and B. Wang (2016). Prospect theory and stock returns: an empirical test. *Review of Financial Studies 29*(11), 3068–3107.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: a review and assessment. *Journal of Economic Perspectives 27*(1), 173–196.

Benartzi, S. and R. H. Thaler (1995). Myopic loss aversion and the equity premium puzzle. *The Quarterly Journal of Economics 110*(1), 73–92.

Bernheim, B. D. (2009). Behavioral welfare economics. *Journal of the European Economic Association 7*(2-3), 267–319.

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

Bernheim, B. D., A. Fradkin, and I. Popov (2015). The welfare economics of default options in 401(k) plans. *American Economic Review 105*(9), 2798–2837.

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

Bernheim, B. D. and A. Rangel (2008). Choice-theoretic foundations for behavioral welfare economics. In A. Caplin and A. Schotter (Eds.), *The Foundations of Positive and Normative Economics*, pp. 155–192. Oxford University Press.

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

Bhargava, S. and G. Loewenstein (2015). Behavioral economics and public policy 102: beyond nudging. *American Economic Review 105*(5), 396–401.

Bleichrodt, H., J. L. Pinto, and P. P. Wakker (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science 47*(11), 1498–1514.

Braga, J. and C. Starmer (2005). Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics 32*(1), 55–89.

Broome, J. (1991). "Utility". *Economics and Philosophy 7*(1), 1–12.

Broome, J. (2009). Why economics needs ethical theory. In K. Basu, S. M. R. Kanbur, and A. Sen (Eds.), *Arguments for a Better World: Essays in Honor of Amartya Sen*, pp. 7–14. Oxford University Press.

Camerer, C. (2008). The case for mindful economics. In A. Caplin and A. Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook*, pp. 43–69. Oxford University Press.

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

Camerer, C. and G. Loewenstein (2004). Behavioural economics: past, present, future. In C. Camerer, G. Loewenstein, and M. Rabin (Eds.), *Advances in Behavioral Economics*, pp. 3–51. Princeton University Press.

Camerer, C., G. Loewenstein, and D. Prelec (2005). Neuroeconomics: how neuroscience can inform economics. *Journal of Economic Literature 43*(1), 9–64.

Caplin, A. and A. Schotter (Eds.) (2008). *The Foundations of Positive and Normative Economics: A Handbook*. Oxford University Press.

Carter, S. and M. McBride (2013). Experienced utility versus decision utility: putting the 'S' in satisfaction. *The Journal of Socio-Economics 42*, 13–23.

Chang, O. H., D. R. Nichols, and J. J. Schultz (1987). Taxpayer attitudes toward tax audit risk. *Journal of Economic Psychology 8*(3), 299–309.

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

Dhami, S. S. (2016). *The Foundations of Behavioral Economic Analysis*. Oxford University Press.

Edwards, K. D. (1996). Prospect theory: a literature review. *International Review of Financial Analysis 5*(1), 19–38.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond "heuristics and biases". *European Review of Social Psychology 2*(1), 83–115.

Gigerenzer, G. and D. G. Goldstein (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review 103*(4), 650–669.

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

Halpern, D. (2015). *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. Allen.

Hands, D. W. (2020). Libertarian paternalism: taking Econs seriously. *International Review of Economics*, 1–23.

Harrison, G. W. and D. Ross (2017). The empirical adequacy of cumulative prospect theory and its implications for normative assessment. *Journal of Economic Methodology 24*(2), 150–165.

Harrison, G. W. and J. T. Swarthout (2016). Cumulative prospect theory in the laboratory: a reconsideration. *Center for the Economic Analysis of Risk Working Paper*.

Hausman, D. M. (2012). *Preference, Value, Choice, and Welfare*. Cambridge University Press.

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

Infante, G., G. Lecouteux, and R. Sugden (2016a). 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.

Infante, G., G. Lecouteux, and R. Sugden (2016b). 'On the Econ within': a reply to Daniel Hausman. *Journal of Economic Methodology 23*(1), 33–37.

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. (2011). *Thinking, Fast and Slow*. Penguin Books.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy 98*(6), 1325–1348.

Kahneman, D., P. Slovic, and A. Tversky (Eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

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 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.

Loewenstein, G. (1988). Frames of mind in intertemporal choice. *Management Science 34*(2), 200–214.

Madrian, B. C. and D. F. Shea (2001). The power of suggestion: inertia in 401(k) participation and savings behavior. *The Quarterly Journal of Economics 116*(4), 1149–1187.

Moscati, I. (2018). *Measuring Utility: From the Marginal Revolution to Behavioral Economics*. Oxford University Press.

Nagatsu, M. (2015). Behavioral economics, history of. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (second ed.), Volume 2, pp. 443–449. Elsevier.

Pinto-Prades, J.-L. and J.-M. Abellan-Perpiñan (2012). When normative and descriptive diverge: how to bridge the difference. *Social Choice and Welfare 38*(4), 569–584.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization 3*(4), 323–343.

Samuelson, W. and R. Zeckhauser (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty 1*(1), 7–59.

Schwartz, B. (2016). *The Paradox of Choice* (revised ed.). HarperCollins.

Sugden, R. (2010). Opportunity as mutual advantage. *Economics and Philosophy 26*(1), 47–68.

Thaler, R. and S. Benartzi (2004). Save more tomorrow™: using behavioral economics to increase employee saving. *Journal of Political Economy 112*(S1), S164–S187.

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

Thaler, R. H. (1987). Anomalies: the January effect. *Journal of Economic Perspectives 1*(1), 197–201.

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

Thaler, R. H. (2018). From cashews to nudges: the evolution of behavioral economics. *American Economic Review 108*(6), 1265–1287.

Thaler, R. H. and H. M. Shefrin (1981). An economic theory of self-control. *Journal of Political Economy 89*(2), 392–406.

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

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

Todd, P. M. and G. Gigerenzer (2012). *Ecological Rationality: Intelligence in the World*. Oxford University Press.

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.

Tversky, A. and D. Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty 5*(4), 297–323.