UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE
ÉCOLE DOCTORALE SCIENCES HUMAINES ET SOCIALES

THÈSE

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE
Discipline : Sciences Économiques

Présentée et soutenue publiquement par

Ivan MITROUCHEV

le 10 novembre 2020

Essays on Normative Behavioural Economics: Methodological and Theoretical Issues

Directeur de thèse : Cyril HÉDOIN, Professeur en Sciences Économiques

Jury

Rapporteurs :

Antoinette BAUJARD, Professeure en Sciences Économiques à l'Université Jean Monnet
Robert SUGDEN, Professeur en Sciences Économiques à University of East Anglia

Président :

Samuel FEREY, Professeur en Sciences Économiques à l'Université de Lorraine

Examinateurs :

Francesco GUALA, Professeur en Économie Politique à Università degli Studi di Milano
Wade HANDS, Professeur en Sciences Économiques à University of Puget Sound

Acknowledgements

As the acknowledgements section is perhaps the part of the thesis where the candidate's creativity is under a bright spotlight, I will try not to disappoint my reader. Yet some conventions are unavoidable. In this matter, it is merely a form of respect to first thank the most important persons who have given value to this academic work.

My first acknowledgement naturally goes to Cyril Hédoin. His pedagogical skills and benevolent attitude were admirable during these four years. His open-mindedness regarding the various methods and objects studied in economics is well aligned with my perception of not restricting myself to a narrow range of knowledge when thinking about a research question. I can only feel lucky to have him as my PhD advisor. Besides, I learnt a lot about preparing a scientific career, which is perhaps as important as making a good thesis.

My gratitude is next addressed to the committee members of the thesis, by whom I feel extremely honoured that they accepted to evaluate my research. The first time I met Antoinette Baujard was at the 4th Economics and Philosophy conference she head-organised in Lyon in 2018. We then became colleagues in 2019 when I was recruited as a temporary teaching and research assistant at the GATE Lyon Saint-Etienne research center. This fourth thesis year has been tremendously enriching for me, talking with her about measuring opportunity and about my future perspectives as a young researcher. Antoinette Baujard has always been benevolent towards me considering my best interests.

Samuel Ferey was one of the two members of my PhD follow-through committee (with Philippe Odou), whom I also met for the first time at the 4th Economics and Philosophy conference in Lyon. I am truly thankful for his remarks and suggestions about how the thesis could be improved, which I hope to have properly taken into account in this final version.

I met Francesco Guala during my first year as a PhD candidate at the 20th Lake Como School of Advanced Studies organised by the European Society for the History of Economic Thought. I am extremely grateful for his warm welcome at the University of Milan during my doctoral mobility. We had weekly meetings discussing Robert Sugden's path-breaking *Community of Advantage*. With the help of Francesco Guala, I learnt a lot in ethics, experimental economics, political philosophy, social choice, theories of justice and welfare economics.

I also met Wade Hands for the first time at the 20th Lake Como School of Advanced Studies. He was at that time my tutor for commenting my very first draft. Spending forty-five minutes with Wade Hands perhaps made me learn much more than I would have done in a full month of reading. I am glad for the interesting discussions, which stimulated me very much to advance in my research topic.

I am grateful towards Robert Sugden for his motivating words about my review essay on his *Community of Advantage*. Reading his book has been extremely beneficial for my understanding of how to propose an alternative approach to the reconciliation problem — a task we engage with Guilhem Lecouteux in Chapter 4 of the thesis. I had the chance to meet Robert Sugden at the conference organised by the Happiness Economics and Interpersonal Relations Association in 2019 where his important book was discussed. From an academic point of view, I think I have exploited as many opportunities as I could during these four thesis years. I imagine Robert Sugden would be happy about this.

This thesis is, properly speaking, not mine. It is the shared work of me and two researchers that I admire not only for their intellectual skills but also for their human qualities. Learning from them confirmed a simple maxim that I have heard several times but could not have the chance to experience until then: working together make you achieve bigger things than you would have achieved on your own.

Working with Guilhem Lecouteux was cognitively demanding. I had a hard time catching up with him on certain things. In fact, I would say that the speed at which my brain processes is surely not calibrated with his own. Yet this collaboration was exciting. Since this thesis is (generally speaking) about the reconciliation problem, reading Guilhem's thesis was compulsory. His work inspired me regarding what a brilliant thesis in economics-and-philosophy should require: references from the history of economic thought, analytical skills, a solid theoretical background and sharp modelling.

Discussing identity with Valerio Buonomo made me change the perspective we economists-philosophers usually have about identity. Thanks to him I understood how complex the literature concerned with what it requires for an individual to persist over time actually is. We are well aware that our study (Chapter 5 of the thesis) is very critical towards the alternative approaches to identity suggested in economics-and-philosophy. In that manner, it may not be well received by the community of economists-philosophers. Conformism is however not my attitude. I am happy to defend the ideas I share with Valerio, and the time has come to defend these ideas.

I also express my gratitude towards my former professors and colleagues who had a great influence on the thesis. I had a very helpful senior colleague at the beginning of my journey: Jérôme Lallement. His thorough knowledge of the history of economic thought helped me a lot in specifying *what* I wanted to say and *how* I wanted to say it. I also thank Annie Cot and Emmanuel Picavet for advising me at the time I graduated to take up the opportunity to start a thesis with Cyril Hédoin, and for telling me that such an opportunity would greatly benefit me. Without doubt they made an accurate prediction.

I am grateful to the people who commented my early drafts at summer schools, namely Guido Erreygers, Wade Hands, Ivan Moscati, George Stefanis, Kate Vredenburgh and Michel Zouboulakis. I enjoyed giving talks at international conferences, workshops, invited seminars and summer schools — sometimes in front of specialists who are now part of my committee. I also thank the persons with whom I had informal discussions at these academic meetings, who significantly helped me in an indirect manner (without them necessarily knowing it). They include Erik Angner, John Davis, Malte Dold, Till Grüne-Yanoff, Leonard Lades, André Lapidus, Ramzi Mabsout, Adam Oliver and Christian

Schubert.

It goes without saying that I thank people from both the REGARDS and GATE research centers for welcoming me as a member of their team. Discussing pretty much any subject in philosophy with Jean-Sébastien Gharbi — from Kant's aesthetics to the foundations of revealed preference theory — helped me a lot to improve my general culture (which is a compulsory requirement for a young researcher in economics-and-philosophy). Discussing normative economics with Sylvain Ferrières, Stéphane Gonzales, Uskali Mäki (who was an invited professor at the GATE research center) and Philippe Solal also highly benefitted me.

I have an obvious thought for my young colleagues Rachid Achbah, Jérémie Bastien, Niels Boissonet, Alexandre Chirat, Francesco Fabbri, Jean-Daniel Houeto, Yao Thibaut Kpegli, Adélaïde de Lastic, Jules Le Lay, David Lowing, George Stefanis and Kevin Techer. Some of them are still in their PhD years, others are postdoctoral fellows, while others made it to assistant professor. I wish them all good luck in what they want to accomplish.

Two things changed my perception about how imperfect my English is, even after having passed a full year in Australia as an exchange student during my Master's degree. First, the amount of corrections Pierre Van Zyl made in my drafts drastically made me think more than once about what I actually know of English. Big thanks to him for his careful reading. Second, *Economical Writing* by McCloskey (2000 [2019]) is a gold mine for non-native speakers writing in English. The good point is that even if you make a bad thesis (which I hope I did not), you spend four years improving both your writing and speaking skills, regardless whether you do your PhD in a foreign or in your own language.

My main extra-academic activities during these four years can without effort be resumed in three: listening to electronic music, playing bullet chess and drinking fruit juices. So let me thank Adam Beyer, Adriatique, Ben Böhmer, Charlotte de Witte, Christian Löffler, Deadmau5, Dubfire, Eric Prydz, Etienne de Crécy, Fakear, Feed Me, Flume, HVOB, Jan Blomqvist, Jon Hopkins, Joris Delacroix, Joseph Capriati, Kiasmos, Klangkarussell, Kollektiv Turmstrasse, Lane 8, Luttrel, Max Cooper, Mind Against, Møme, No Mana, Oliver Koletzki, Porter Robinson, Ryan Davis, Sama, San Holo, Shingo Nakamura, Stephan Bodzin, Tale of Us, Transcode, Yotto, the Afterlife, Anjunadeep, Drumcode and Mau5trap labels for providing me with evasion, enthusiasm and peace of mind. Let me also thank Booba, Dinos, Jazzy Bazz, Kaaris, Kekra, Lomepal, Nekfeu, Népal, Niska, Orelsan and PNL for providing me with other kinds of emotions.

Since my PhD advisor will obviously read this, the hours spent playing bullet chess on Lichess.org during these four years will not be disclosed (to whom it may concern, now that this thesis is done my challenge still applies). Thanks to Mohamed Zeroual and Antonio Radic (a.k.a. agadmator) for being such great trainers. My style of playing chess (tactical blunders, dubious sacrifices, pattern 'nonrecognition') and my taste for adventure made me observe that my decision-making attitude is quite risk-seeking. From my own viewpoint I could add that I only have a normal attitude towards life, which would otherwise be quite boring.

I end up with a few words addressed to people who had less influence in this thesis but to whom I am emotionally attached because they are my friends. In addition to chess,

electronic music and fruit juices I had a (bit of a) social life. I am particularly glad to have had Enzo Camilla, Umar Green and Mario Ramirez as flatmates during my staying in Milan. Playing FIFA with Enzo, partying with Umar and watching movies with Mario was fun. A special thought goes to my good friend in Reims Célia Prévost. I am glad she has now the life she desires and deserves.

Explaining my thesis for Lucille Belmonte while we were at McDonald's helped me much in advancing in my research. Lucille and I spent hours talking about issues related to the commensurability of sensations (Chapter 2 of the thesis) and about the different normative criteria offered in the literature (Chapter 3 of the thesis). I was always surprised by her lucidity. She asks naive-disturbing questions that would make specialists embarrassed about what they are paid to do every day.

The person who gave me the most psychological support during these four years is my old friend Anaïs Delotterie, whom I tremendously admire. I say the same for Sarah Delotterie. Tragedies ought to be overcome, but the gap between saying it and doing it is abyssal. My last thoughts go to Clément Casali, Maxence Pichon and Cédric Polano, who perhaps know me better than anyone else. I hope that with this thesis they will be proud of me.

I emphasise that I wrote fragments of these acknowledgements a year and a half before the submission of the thesis in order to avoid the *availability bias*: putting more emotional weight into what happens at the end of the fourth thesis year and then partially neglecting the whole process to which I have arrived. Who said the acknowledgements had to be written at the end? I believe a proper reconstruction of emotional states requires one to give importance to what happened during *all* four years. This allows me to thank the most important people for these years that have been so intellectually stimulating.

Considering myself as a disciple of David Hume, the degree of existence of something depends on one's emotional involvement. I can say that this thesis had a very high degree of existence for me, but that its perceived value will of course depend on each one's idiosyncratic reference point. I hope to make its perceived value as high as possible.

Contents

List of Figures

List of Tables

GENERAL INTRODUCTION

You go shopping at the supermarket. You have the choice between plenty of goods: chocolate cakes, apples, boxes of pasta, ketchups, juices, etc. At the moment you make a decision to buy one item rather than another, one may wonder whether you made the 'right' choice. As nobody except you is supposed to know whether you actually made the 'right' choice, economists have for a long time assumed that whatever people choose is best for them. Why would economists really think so? People may actually not like the ketchup they thought was tasty, or they may make decisions they may later regret (e.g. being tempted by the nice chocolate cake). The simple answer economists have had so far is that nobody has had 'scientifically' good enough arguments to question the contrary (except perhaps philosophers). Importantly, assuming the contrary would violate a liberal principle that is fundamental not only for the standard tradition of economic theory but also for our modern societies: no one can be a better judge of what is best for oneself than oneself.

Thus, in standard economics policymakers and society respect *observed choices* because it is assumed that they reflect a considered judgement of what makes consumers better off. With the large amount of empirical evidence that consumers — that I will more generally call 'individuals' in this thesis — do not have preferences which conform to the rational norms of standard economics, many behavioural economists however think that this evidence may provide a 'scientific' background for the fact that individuals often do not choose according to their own interests.

'*Is it so?*' is the question I address in this thesis. Is the fact of regretting your choice, being not informed about what you bought, or being tempted by the nice chocolate cake necessarily makes your choice 'wrong'? Nothing is really sure, simply because empirical evidence of incoherent preferences is far from being enough to justify that one is making a mistake. In addition, we need to question the underlying assumptions of behavioural economists about what makes individuals better off. In this thesis I question and discuss those assumptions, where they may come from, what arguments are involved to justify them or refute them, and what new directions we can propose to solve some of the important methodological and theoretical issues related to the normative implications of behavioural economics.

Normative Behavioural Economics

Before making the research question and the outline of the thesis explicit, it is first necessary to explain what normative behavioural economics is about. Economists are now quite familiar with what behavioural economics is (the field of introducing a psychological background in order to enrich economic theory), and certainly even more familiar with the well-known positive-normative distinction in economics. So can we speak of behavioural

economics as being also separable into two branches: 'positive' behavioural economics, on the one hand, and 'normative' behavioural economics, on the other hand? Briefly speaking, yes. Behavioural economics refers in fact to 'positive' behavioural economics. But since 'positive' adds no informative value to characterise behavioural economics, the specification of 'positive' is redundant. Instead, 'positive' behavioural economics is specified when we need to contrast it with the other branch of behavioural economics that uses observations of decision-making for normative analysis: 'normative' behavioural economics.

The terminology of 'normative behavioural economics' is far from being new. It is originally used by Berg (2003), who defends a methodological pluralism for normative analysis. The author criticises the overuse of the *homo-economicus* (that I will call Econ in the thesis) as a normative benchmark, which, according to him, in many cases does not provide a connection between behavioural hypotheses and policy prescription (p. 424). It is in Berg's (2003) spirit — i.e. questioning the assumptions of behavioural economists concerned with normative analysis — that I retain the terminology of 'normative behavioural economics' as the subject matter of this thesis.

Normative behavioural economics is also the terminology used by Ogaki and Tanaka (2017 [Chap. 11]). The authors present the field of normative behavioural economics by starting at the distinction between positive and normative economics. By introducing behavioural economics as a branch of economics, they note that behavioural economics has also two subbranches: 'positive behavioral economics' and 'normative behavioral economics' (p. 185).

Substantially, we can define normative behavioural economics as the need for rules that can be used to assess the desirability of public policies. But to contrast it with 'standard' normative economics, we shall add *...with a special focus on the psychology of individuals*. To put it differently, normative behavioural economics is the field which deals with any normative concern that derives from behavioural economics. Those normative concerns can be the evaluation of individuals' states of affairs or policy recommendation/prescription based on given normative criteria. Henceforth, I will use the general term 'normative analysis' to refer to those normative concerns.

Now what does it mean that normative analysis 'derives' from behavioural economics? Such positive-to-normative relationship is about the interpretation of *preference*, the primitive concept on which economic theory is originally grounded. In the standard economic approach of decision-making, individuals are simply assumed to meet the definition of rationality. In this sense, the satisfaction of preference entails *both* positive and normative economics: there is no distinction between how individuals *do* and *ought to* choose. Behavioural economics however challenges the positive validity of the norms of rational choice by providing numerous cases where individuals have preferences which violate the norms of rational choice theory. Here is a list of some of them.

- *Limited attention*. A substantial amount of empirical evidence supports that individuals are limited in their ability to process information and to perform multiple tasks simultaneously (Kahneman 1973). When individuals have limited attention, it requires them to allocate their cognitive resources across tasks, so that attention

spent on one task consequently reduces attention available for other tasks. This finding goes against the hypothesis of rational choice, according to which the rational individual has no deficiency of this sort. Consequently, some behavioural economists came to the intuitive conclusion that had the decision maker been perfectly informed (or had she disposed of full cognitive abilities), she would have perfectly known what to choose among available alternatives.

- *Non-Bayesian updating*. Models of rational choice typically assume that individuals update available information according to Bayes' rule. But there is a consequent body of empirical evidence that does not support this assumption (Tversky and Kahneman 1973, 1974). The implication is that not only non-Bayesian updating may falsify some market phenomena such as the efficient market hypothesis, but it may also make investors deviating from optimal decisions on stock markets (Thaler 1987).

- *Time-inconsistency*. In the standard model of rational choice, the decision maker has the same preferences about future plans at different points in time. The assumption of time consistency is captured by the exponential (δ) discounting model (Samuelson 1937). It is however often considered that hyperbolic or quasi-hyperbolic (β, δ) discounting models are empirically more consistent with observed preference over time (Laibson 1997; O'Donoghue and Rabin 1999). These models are particularly well appropriate to explain self-control problems (Thaler and Shefrin 1981), i.e. the failure of managing one's willpower like being tempted by the nice chocolate cake.

- *Context-dependence*. There is now a large body of empirical evidence which supports the idea that individuals have different preferences according to the context in which they choose, even if the available alternatives are actually identical (Tversky and Kahneman 1981, 1986). This observation is in conflict with the context-independency assumption of rational choice, according to which adding an irrelevant alternative should not change one's preference (Allais 1953) or the order in which alternatives are presented should not depend on the manner in which these alternatives are described (Kahneman and Tversky 1984). One implication is that individuals' choice may be influenced by *framing*, which eventually may not lead them to choose what makes them better off.

- *Loss aversion*. The principle that losses loom larger than corresponding gains is supported by a fair amount of empirical evidence (Kahneman, Knetsch, and Thaler 1991; Tversky and Kahneman 1991). This psychological phenomenon may imply the known *endowment effect* (Thaler 1980) and *status quo* bias (Samuelson and Zeckhauser 1988). Again, these observations are not in accord with the principle of rational choice, according to which preferences do not depend on current assets. The implication is that an external observer may judge this behaviour to be 'irrational' if he judges there is no particular reason for the individual to have this aspect of her psychology.

These inconsistencies in behaviour not only concern the long tradition of positive economics, which aims at describing, explaining and predicting market behaviour based on the Econ abstraction. They also concern the standard tradition of normative economics, which takes the satisfaction of coherent preferences as the central criterion for normative analysis. Standard normative economics (which archetype is standard *welfare* economics)

assumes that individuals' preferences are coherent — that is, consistent with the standard axioms of rational choice, stable and context-independent — and that the satisfaction of these coherent preferences is the normative criterion.[1]

Yet an important issue for normative analysis is that if empirical studies systematically show that individuals' preferences are incoherent, their observed preferences cannot reasonably indicate their well-being. This is particularly true in circumstances where we have doubts that individuals are able to meet the standard assumptions of rational choice, e.g. when the decision involves risk or uncertainty. The term 'observed preference' has here a normative significance. If the chosen alternative is always also the preferred one, preference-satisfaction should be considered as a decent indicator of one's well-being. However, because of these violations of rational choice (among others) the standard preference-satisfaction assumption has recently been disputed by many behavioural economists, who uphold several alternative approaches to normative analysis. Those can essentially be resumed by the following three classes.

- *Experienced utility measurement*. If one judges that observed preference may not always be a good indicator of individual well-being, one may find it useful to make a theoretical distinction between *decision utility* (the weight of a decision inferred from observed choice) and *experienced utility* (the hedonic state experienced in doing or choosing something). The idea is to take only experienced utility instead of decision utility as the proper criterion for normative analysis. This approach has been developed and supported by a handful of researchers at the beginning of the 1990s, who originally disputed the idea that observed preference is a good indicator of well-being (Kahneman and Snell 1990, 1992; Kahneman and Varey 1991; Varey and Kahneman 1992). Their project was then concretised in the theory of bringing happiness measurement 'back to Bentham' (Kahneman, Wakker, and Sarin 1997).

- *Behavioural welfare economics*. Another alternative is to consider only preferences (or choices) undistorted by cognitive biases to be the proper criterion for evaluating individuals' states of affairs. In the line of the old tradition in standard welfare economics — which takes Pareto efficiency and utility as its sole basis for normative analysis — behavioural welfare economics is the mainstream approach in normative behavioural economics. 'Behavioural welfare economics' is a term that originated in Bernheim and Rangel (2007) and is taken as a subject of its own in Dhami (2016 [Part. 8]). The field is presented as any study which deals with the welfare implications of behavioural economics. The aim of behavioural welfare economics is not to *help* individuals making better decisions. Its main purpose is to extend the standard framework of welfare economics with the introduction of cognitive biases. But whether paternalistic policies are desirable is yet another question.

  - *Behavioural Paternalism*. This last question concerns behavioural paternalism. Such field is to be defined as the branch of policy analysis aiming at exploiting

[1] The standard axioms of rational choice are completeness [$\forall x, y \in X, x \succcurlyeq y$ or $y \succcurlyeq x$] and transitivity [$\forall x, y, z \in X, x \succcurlyeq y$ and $y \succcurlyeq z \implies x \succcurlyeq z$]. Stability means that the individual's preferences are stable over the period for which the external observer observes her choice behaviour. Context-independency means that preferences should not vary depending on the context of the choice (or in the manner the choice is presented). Unless specified, the combination of these axioms is the only way I will use the term 'coherent preference' throughout the thesis.

individuals' cognitive biases in order to increase their well-being with no (or minor) harm on either other individuals or themselves. Behavioural paternalism can take few subtle forms. Two of them are widely known.

  * *Asymmetric paternalism*. This approach is advocated by Camerer et al. (2003) in a manifesto about the usefulness of behavioural economics for policy analysis. The aim of the authors is to show that cost-benefit analysis can be extended to irrational behaviour. They argue that it is possible to increase the benefits of those who make errors with little or no harm on those who are fully rational.
  * *Libertarian paternalism*. This approach is advocated by Thaler and Sunstein (2003, 2009) and had a huge impact on the public sphere with the institution of Behavioural Insight Units (mostly known as 'Nudge Units') all over the world (Halpern 2015). Libertarian paternalism aims to improve individual decision-making with no (or minor) limitation on one's freedom to choose among available alternatives.

- *Institutional arrangement*. Another approach to normative behavioural economics is simply not to account for happiness nor incoherent preferences, but instead to promote institutional arrangements so that individuals are better able to satisfy their preferences, *whatever they are* (i.e. coherent or not). This approach is suggested by Sugden (2004) and given an extensive support in Sugden (2018a). The essence of institutional arrangement is to provide individuals with the opportunity to choose from the available goods in the economy by letting them be their own judge of what makes them better off. In this matter, the approach is anti-welfarist and anti-paternalistic. It takes opportunity instead of utility as its informational basis, and sticks to the consumer sovereignty principle of standard economic theory. According to this principle, nobody is supposed to judge whether the nice chocolate cake makes you worse off (had you been tempted to buy it).

With this succinct overview of normative behavioural economics, one can say that albeit the field is relatively new (approximately thirty years old), it already offers a rich set of propositions to normative analysis. Indeed, libertarian paternalism — the most influential and discussed approach in the public sphere — is only an element of behavioural paternalism, which belongs to the set of behavioural welfare economics, which itself belongs to the set of normative behavioural economics.

The scope of the thesis entails the broad field of normative behavioural economics and not libertarian paternalism in particular. One reason is that although being mainstream, libertarian paternalism (and more generally behavioural welfare economics) is far from being the only problematic approach in normative behavioural economics from both methodological and theoretical points of view. There is also no particular reason to focus on one normative approach over another, simply because *all of them* are relevant to the ambitious challenge of making normative economics consistent with behavioural economics.[2]

[2]This challenge is coined as the 'reconciliation problem' by McQuillin and Sugden (2012). Fundamentally, there is no distinction between normative behavioural economics and the reconciliation problem. The only difference may be that normative behavioural economics refers to the normative *branch* of behavioural economics, while the reconciliation problem refers to the *problem* of reconciling normative and behavioural economics.

Research Question

So what is at stake in normative behavioural economics? From a general point of view, there seems to be a tension between descriptive and normative decision-making when the two do not converge on joint notions such as the satisfaction of coherent preference as the normative criterion. For almost a century, the ordinalist tradition considered mental states (and more generally psychology) to be 'out' of economics (Hands 2010). However, if we aim to do normative analysis it seems difficult to avoid focusing on individual psychology. This is because individuals seek to achieve personal goals they construct from their ethical values, which are not only intrinsically located in their *mental states* but also located in external social phenomena represented by *institutional arrangements*.

The main challenge of normative behavioural economics (as I see it) is thus to improve the way to evaluate/recommend/prescribe public policies based (i) on the available empirical evidence on human behaviour *but also* (ii) on the various philosophical problems regarding what makes individuals better off. The thesis proposes some answers to this important enquiry by taking the second point seriously.

The economics-and-philosophy literature has already been extensively concerned with the problems of normative behavioural economics in many ways. Most of this literature particularly focuses on behavioural welfare economics, more particularly on behavioural paternalism, and even more particularly on libertarian paternalism.[3] This thesis contributes to this related literature. It provides a conceptual analysis of the methodological and theoretical issues of normative behavioural economics in the following respects.

Outline of the Thesis

Chapter 1 is a historical analysis of the influential role prospect theory may have had on behavioural welfare economics. It is mainly addressed to the community of economists-philosophers, whose aim is to clarify some methodological issues of behavioural welfare economics. Chapter 2 is a philosophical assessment of measuring experienced utility. It is addressed to economists and policymakers who are interested in measures of objective happiness alternative to experienced utility. Chapter 3 is a survey of the reconciliation problem since there is (to my knowledge) no existing synthesising work which makes an extensive taxonomy of the main normative criteria offered in normative behavioural economics and their associated methodological and theoretical issues. This chapter is addressed to the community of economists interested in 'solving' the reconciliation problem. Chapter 4 and 5 engage in new directions for normative behavioural economics.

[3]This non-exhaustive literature includes Hausman (2008, 2012, 2016, 2018), Sugden (2008, 2015, 2017b, 2018b), Rizzo and Whitman (2009, 2018, 2019), Berg and Gigerenzer (2010), Welch and Hausman (2010), Ferey (2011), Grüne-Yanoff (2012, 2016, 2018), Rebonato (2012), Baujard (2015), Guala and Mittone (2015), Gigerenzer (2015), Hédoin (2015, 2017), Marciano (2015), Nagatsu (2015b), Whitman and Rizzo (2015), Grüne-Yanoff and Hertwig (2016), Infante, Lecouteux and Sugden (2016a, 2016b), Davis (2018), Dold (2018), Dold and Schubert (2018) and Hands (2020). In addition, two PhD theses about the normative implications of behavioural economics have been defended in the past few years (Lecouteux 2015b; Dold 2017). To my knowledge, this is the third (and probably not the last) PhD thesis about this promising area of research.

Chapter 4 is a joint work with Guilhem Lecouteux. We develop a normative approach which accounts for context-dependent preferences. Chapter 5 is a joint work with Valerio Buonomo. We propose an ontological framework of personal persistence in order to better understand the ethical problem of identity associated with the assumption of multiple selves in behavioural welfare economics. These last two chapters are addressed to the community of economists and economists-philosophers interested in alternative approaches to normative behavioural economics.

Chapter 1

It is common to see normative behavioural economics as a field that emerged (approximatively) between the 1990s and the 2000s with the revival of the Benthamian-Edgeworthian measurement of happiness (Kahneman, Wakker, and Sarin 1997) and with behavioural paternalism (Camerer et al. 2003; Thaler and Sunstein 2003). The aim of this first chapter is to question whether the concerns behavioural economists had towards normative analysis were not older than the 1990s, and if they actually were, how they could bring new insights into some of the known methodological problems of behavioural welfare economics. My aim is to investigate the early normative concerns Kahneman and Tversky had in their first proposition of prospect theory (1979, 1981, 1986) and to argue that those early references to normative analysis are informative to the methodology of behavioural welfare economics in at least three ways. First, they provide an explanation of why the heuristics-and-biases program tends to consider deviations from rationality as *biases*. Second, they provide an explanation of the practical usefulness of distinguishing descriptive and normative decision-making as two separate enterprises. Third, investigating the early stage of prospect theory helps to clarify the difficulties associated with the elicitation of individuals' true preferences. The overall argument of this chapter is that contrary to the common view that the heuristics-and-biases program neglected any normative concern, I here provide a nuanced view. Founders of prospect theory *did* share some early concerns about the normative implications of their theory, and the program of behavioural welfare economics is more likely to be seen as a natural extension of the heuristics-and-biases program rather than a new research program *per se*.

Chapter 2

This chapter constitutes an analytical assessment of the whole program of measuring experienced utility. In a recent interview, Daniel Kahneman explicitly acknowledged that measuring happiness in terms of hedonic maximisation may actually be misleading. My goals are (i) to make a census of the important reasons that measuring experienced utility seems to be a dead end, and (ii) to suggest an alternative direction to objective happiness measurement. I first review the literature of twenty years of research on measuring experienced utility. I then consider several reasons that Benthamian hedonism may be problematic for public policy. Then, I propose a philosophical discussion of all the axioms of experienced utility measurement by pointing out most of their theoretical issues. Finally, I show that maximising individuals' *moment* utilities (what is experienced here and now) appears to be based on a misconception of happiness that economists and policymakers have good reason to stay away from. The bottom line is that if economists and policymakers seek to improve their understanding of measuring objective happiness, they should better cut off with Bentham's (1780 [2007]) hedonistic reductionism as

their ethical benchmark. Instead, I argue that a more convincing approach to objective happiness measurement can be found in Aristotle's (-350 [2009]) ethics. Contrary to Bentham's hedonism, Aristotle's eudaimonsim considers pleasure to be only a *by-product* of happiness but not something that *constitutes* happiness. This conception of objective happiness is more appealing for economists and policymakers because (i) it fits better with the range of what public policy applies to, (ii) it is better aligned with the methodological and theoretical issues associated with experienced utility measurement, and (iii) it is not in contradiction with recent empirical evidence that shows no discrepancy between experienced utility and decision utility. Aristotle's conception of happiness is in fact closely related to Daniel Kahneman's recent statement. According to this statement, objective happiness is to be defined in terms of social relationships and what individuals remember of their experiences rather than what is experienced here and now.

Chapter 3

This chapter is an attempt to find a consensus on how the reconciliation problem can be best tackled. As initially introduced by McQuillin and Sugden (2012), there is neither an existing literature review of the reconciliation problem nor a consensus about how the problem is best tackled. This chapter aims at filling both of these gaps. I first categorise four classes of normative criteria that are now well developed in normative behavioural economics: *experienced utility*, *true preference*, *choice-basis* and *opportunity*. The experienced utility criterion breaks with the concept of preference and aims at maximising individuals' experiences of pleasure (or minimising their experiences of pain). The true preference and choice-based criteria take the satisfaction of individuals' preference/choice undistorted by cognitive biases to be the proper criterion for normative analysis. The opportunity criterion instead breaks with the assumption of rationality as the normative benchmark and aims at enhancing individuals' opportunities to choose from. My goal is to critically examine the strengths and weaknesses of each of these normative criteria so that new perspectives of research can be suggested regarding the methodological and theoretical difficulties encountered by each normative criterion. To compare these normative criteria, I propose a simple framework based on what I judge to be the essential question of the reconciliation problem: *what is a good normative criterion?* Accordingly, I propose three fundamental requirements that a good normative criterion should satisfy. First, a normative criterion should apply to a wide range of choice situations (what I call the *general* requirement). Second, a normative criterion should capture the many different aspects of life that individuals can find valuable (what I call the *ethical* requirement). Third, a normative criterion should measure individuals' states of affairs using a relatively consensual measure of what makes individuals better off (what I call the *practical* requirement). This simple framework allows me to evaluate whether each normative criterion fulfils these three requirements. The result is however that none of them satisfy them all. This leads me to suggest avenues of future research on promoting alternative normative criteria that could potentially better satisfy these three requirements. These alternative normative criteria are the *virtue ethics* criterion and the *meaning* criterion.

Chapter 4

In this chapter we develop a social choice approach of normative behavioural economics that accounts for individuals' context-dependent preferences. Our goal is to highlight that most of the debate around the reconciliation problem is fundamentally similar in nature to questions in social choice theory about how to aggregate individual preferences. While social choice is about preference *aggregation*, the reconciliation problem is about preference *integration*. This means that, on the one hand, social choice typically starts from the individual and is then interested in the social evaluation of outcomes/institutions. On the other hand, the reconciliation problem implicitly starts from the multiple selves assumption of the individual and is then interested in how to integrate the preferences of the different selves into a single individual. The other essential question of the reconciliation problem we tackle here is '*to whom normative economics should be addressed?*' (Sugden 2018a). We focus on the standpoint from which the normative preferences of the individual should be defined, just as social choice theorists may wonder about the correct standpoint from which a social welfare function (or more generally, any normative criterion) should be defined. Two approaches are currently offered in the literature. The 'third-person' standpoint is the approach of *experienced utility measurement* and *behavioural welfare economics*, where a social planner aggregates individuals' preferences. The 'first-person' standpoint is the approach of *institutional arrangement*, where it is up to individuals to define what their own good is and then to engage in mutually beneficial exchanges. The alternative approach we propose is the 'second-person' perspective inspired from contractualism in social choice. In this approach, normative analysis focuses on the *process* of preference integration, i.e. the process by which individuals' multiple selves start with conflicting preferences and end up with their own preferences (that are not necessarily coherent). That is, the normatively relevant inputs are neither only the preferences of the individual (like in the 'first-person' standpoint) nor only the preferences of the contradicting selves (like in the 'third-person' standpoint). Instead, it is the *process* through which the individual integrates the preferences of her different selves into her own (final) preferences. The originality of our approach is that instead of proposing a single acceptable context which is either (i) exogenous to the normative representations of individuals (*view from nowhere*), or which (ii) only accounts for their behaviour but not their internal processes that lead them to their own preferences (*view from somewhere*), we propose that normative analysis can focus on how individuals may confront the views from different contexts so that they can form their own normative judgements (what we call the *view from 'manywhere'*).

Chapter 5

In the last chapter we explore the ethical problems of multiple selves from an ontological viewpoint. Multiple selves is the conventional assumption in behavioural welfare economics for modelling individual well-being over time. One concerning issue is if preferences change over time, which of the many possible individual's preferences are normatively relevant? To palliate this issue, some theoretical and philosophical contributions argue for the relevance of the unified self assumption against the common representation of the dualistic concept of the individual in behavioural welfare economics.[4] Albeit ap-

[4]These contributions include Sugden (2004), Ferey (2011), Lecouteux (2015a), Hédoin (2015, 2017), Gallois and Hédoin (2017) and Dold and Schubert (2018).

pealing, this approach of the unified self remains however unclear from a philosophical viewpoint because it involves ontological debates about the status of what makes an individual *persist* through time — a vocabulary used in analytical philosophy for what is often referred to as 'agency' in economics. Our aim is to clarify the philosophical difficulties of engaging through this unifying view of the self, what has so far been undeveloped. We claim that for those alternative assumptions of the unified self to hold, an important challenge is to determine what makes an individual persist through time. We argue that ethical questions related to modelling temporal selves can be informed by the ontological question of personal persistence. Based on a simple analytical framework, we review the various theories of personal persistence offered in analytic philosophy and provide philosophical arguments about why the narrative view of the self — which is the dominant unified theory of personal persistence endorsed in the critical literature of behavioural welfare economics — is unwarranted. Our main result is that the assumption of the unified self is philosophically as problematic as the assumption of multiple selves. Our study implies new paths of research regarding the exploration of personal persistence for normative economics — a domain that is yet unknown to the literature of identity-and-economics.

Reader's Guide

To summarise, this thesis aims at providing some answers to the following questions. Is normative behavioural economics that recent (and why does it matter)? (Chapter 1). What are the methodological and theoretical issues of the experienced utility criterion, and how can one overcome those issues from an ethical viewpoint? (Chapter 2). How can one reach a consensus on solving the reconciliation problem? (Chapter 3). What normative standpoint should one take when individuals' preferences are context-dependent, or to whom should normative economics be addressed? (Chapter 4). What is philosophically problematic with the alternative unified self assumption in normative economics? (Chapter 5).

The order of these chapters constitutes what I judge to be a consistent narrative to the subject matter of the thesis. That is, Chapter 1 provides a brief historical analysis of normative behavioural economics. Then, Chapter 2 and Chapter 3 provide extensive literature reviews and critical assessments of the main approaches in normative behavioural economics. Finally, Chapter 4 and Chapter 5 propose new directions for normative behavioural economics.

The reader must however feel free to read the chapters randomly as they are written in a way that they do not require any knowledge from the other chapters. This also allows readers from different backgrounds to focus on what they are more interested in. For example, researchers in happiness measurement may find more interest in Chapter 2, researchers in social choice may find more interest in Chapter 4, while researchers in identity may find more interest in Chapter 5.

What This Thesis Is Not About

We are now ready to begin studying the various methodological and theoretical problems of normative behavioural economics. And without further due, the impatient reader may already skip to Chapter 1 (or any other chapter). However, since there are obviously many aspects that could not have been studied in this thesis for numerous reasons — principally literature positioning and time constraints — I shall end up this general introduction by informing the reader of some related content that is outside the scope of the thesis.

'Positive' Behavioural Economics

The first and perhaps most important awareness that should be given is that I provide no defence of the empirical adequacy of the cognitive biases documented in behavioural economics such as limited attention, non-Bayesian updating, time-inconsistency, framing and loss aversion. In other words, the thesis does not address in any way questions about the empirical adequacy of the cognitive biases in the literature. Instead, I offer a critical analysis of the *normative implications* of these cognitive biases (among others).

That is, assuming that behavioural economics is right in characterising those behaviours as demonstrating inconsistencies of choice, what does that imply about the normative criteria proposed in normative behavioural economics, and more generally about the way to do normative economics? I then apologise if I here repeat myself. This thesis is not about 'positive' behavioural economics but about 'normative' behavioural economics.

Obviously, since 'normative' behavioural economics is nurtured by 'positive' behavioural economics (the former only exists because of the latter), some references to inconsistent behaviour (such as the ones previously listed) are necessary. The point is that I do not debate whether any form of incoherent preference is empirically adequate. Instead, I simply take the observed phenomenon of incoherent preference as *given*.

The Positive-Normative Relationship *in Itself*

Although this thesis is very much embedded in the economics-and-philosophy literature concerned with the 'positive-normative' relationship in behavioural economics, no philosophical investigation about how one can pass from an *is* to an *ought*, nor from judgements of *fact* to judgements of *value* are undertaken. How Hume's law can be bypassed is subject to a huge investigation in metaethics, which for evident reasons could not be part of the present work.

However, with the decline of the fact/value dichotomy (Putnam 2002), Hume's law is not as appealing as it used to be. Indeed, the arguments that judgements of fact are *stricto sensu* different from judgements of value have been subject to insightful criticisms over the past few years.[5] Even though I do not undertake any philosophical investigation of the relationship between positive and normative economics, a clarification of the two categories is obviously useful.

[5] See Dasgupta (2005), Kincaid, Dupré, and Wylie (2007), Putnam and Walsh (2011) and Hands (2012).

A statement is positive when it involves a judgement of *fact*, which can be either refuted or falsified by empirical evidence. Positive judgements can take at least three forms: descriptive, explanatory or predictive. Descriptive statements provide an information of the observed phenomenon. For example, 'the 2019 GDP of France has increased by 1.2%' or 'an individual prefers a sure gain of 240$ to the risky gamble of (0.25, 1000$)' are descriptive statements.

Contrary to explanatory statements, descriptive statements do not provide a *reason* for the observed phenomenon. Instead, statements such as 'the 2019 GDP of France has increased by 1.2% because the economy was not yet affected by a pandemic' or 'an individual prefers a sure gain of 240$ to the risky gamble of (0.25, 1000$) because she is risk-averse' are explanatory statements.

Predictive statements relate to an uncertain knowledge based on the information we have about descriptive and explanatory statements. It is a likely state of the world to occur in the future based on empirical evidence. For example, 'the 2020 GDP of France will decrease by more than 10% because the economy is currently affected by a pandemic' or 'an individual will prefer a sure gain of 240$ to the risky gamble of (0.25, 1000$) because we infer from repeated experiments that individuals are typically risk-averse' are predictive statements.

Normative statements, on the other hand, can also take at least three forms: evaluative, recommendative and prescriptive. An evaluative statement establishes a rule that one outcome (or state of the world) is better than another. For example, 'the 2019 GDP of France is better than the predicted 2020 GDP of France' is an evaluative statement.

But to evaluate does not necessarily imply to recommend nor to prescribe a policy based on this evaluation. Instead, a statement such as 'the 2020 GDP of France *should* be higher than the 2019 GDP of France' provides a picture of the desirable world conveyed by its announcer (the economist), which can be either recommendative or prescriptive. There is now a difference in degree between recommending and prescribing.

Recommending is 'authoritatively weaker' than prescribing. That is, shall the statement 'the 2020 GDP of France should be higher than the 2019 GDP of France' be taken as a recommendation, its announcer is *advising* the public on the desirable state of the world but takes no position on whether her statement will actually be followed or not. Prescribing is 'authoritatively stronger'. Shall the statement 'the 2020 GDP of France should be higher than the 2019 GDP of France' be taken as a prescription, its announcer is *ordering* or *commanding* what is desirable for the public (of course, this does not exclude the possibility that the public can rebel against the order).[6]

[6]One may consider 'recommending' and 'prescribing' to be synonyms, which is in my view misleading because it fails to nuance between *advising*, on the one hand, and *commanding/ordering*, on the other hand. The differences between descriptive, explanatory and predictive statements, on the one hand, and between evaluative, recommendative and prescriptive statements, on the other hand, are here only useful to bring precision to the meaning of positive and normative statements. They do not constitute here a part of study *per se*.

Market Failure

I have already stated that I will use the term 'normative analysis' to refer to the evaluation of individuals' states of affairs and policy prescription/recommendation based on given criteria. However, normative analysis obviously aims to provide answers to a wide panel of questions such as how a social welfare function should be constructed, how markets should be regulated, or whether public goods should be supplied by government, and if so, which goods in which quantities.

In order to restrict the scope of the thesis, I take 'normative analysis' in a narrow sense. That is, I am only concerned with the assessment of *individual* states of affairs, not *social* states of affairs. The reader must not be surprised if I say nothing about situations where an individual is affected by the choice of another individual. Being silent about social interactions is typical of most welfare economics and also behavioural welfare economics. This thesis is however not concerned with the literature on market failure, which is yet another vast domain of normative economics to be treated separately.

According to some economists, behavioural economics constitutes a big challenge for normative economics in the way that it introduces new forms of market inefficiency at the individual level (Beshears et al. 2008; Madrian 2014; Chetty 2015). In standard welfare economics, the typical taxonomy of public goods, externalities, information asymmetries and market power is a source of market failure (or economic inefficiency) and justifies government regulation/intervention through taxation, regulating output and mandating information disclosure in order to increase or (at best) re-establish market efficiency.

The new source of market inefficiency added by behavioural economists is the individuals' cognitive biases such as the ones previously listed (limited attention, non-Bayesian updating, time-inconsistency, etc.). With the development of behavioural economics, the focus of policy tools is not on the shifting of prices but on the shifting of behaviours towards optimal choices. That is to say, the focus has moved from an *interpersonal* to an *intrapersonal* level of normative analysis, i.e. a shift from *externality* to *internality*.[7]

Bracketing out market failure (especially externalities) does not mean I consider social evaluation to be less important than individual evaluation. Quite the contrary: if public policy is the *telos* of normative economics, new forms of public policies that derive from behavioural economics (called 'behavioural' public policies) are directly concerned with market failure. In this matter, it is perhaps of greater concern to study the effect of behavioural paternalism when externalities are involved (Guala and Mittone 2015).

It is however also useful to understand rules at the individual level. In this thesis, I exclusively focus on the latter. We can thus see normative behavioural economics as having two subbranches: (i) studying the rules for assessing individuals' states of affairs — which is the aim of this thesis — and (ii) studying at the social level the implications of market failure (e.g. externalities) — which is outside the scope of this thesis.

[7]The term 'internality' is first introduced by Herrnstein et al. (1993) and is defined as the failure individuals have of maximising their own utility.

Alternative to the Heuristics-and-Biases Program

While there are many aspects to behavioural economics, the thesis focuses exclusively on the heuristics-and-biases program (Kahneman, Slovic, and Tversky 1982; Kahneman and Tversky 2000) because it is by far the most influential and because it is the program most relevant to the literature of normative behavioural economics. It is however clear that the most influential other research program in the contemporary literature on behavioural economics is the fast-and-frugal-heuristics program (Gigerenzer and Selten 2001; Todd and Gigerenzer 2012).

To give a brief overview of these two 'rival' programs, one major disagreement between tenants of the fast-and-frugal-heuristics program and tenants of the heuristics-and-biases program is about the interpretation of rationality as a positive concept.[8] Basically, the former blame the latter for providing an unrealistic account of how individuals make decisions. Tenants of the fast-and-frugal heuristics program argue that it makes no sense to give importance only to cognitive limitations of individuals and not to the environment in which these individuals make decisions.

The common metaphor they use is a pair of scissors inspired from Simon's (1956) idea that the environment possesses properties that permit further simplification in choice mechanisms (p. 129). One blade represents the 'cognitive limitations' of actual humans (what tenants of the heuristics-and-biases program take into account), while the other represents the 'structure of the environment' (what tenants of the fast-and-frugal-heuristics program blame the heuristics-and-biases program for not taking into account).

Based on their very distinctive interpretation of the concept of 'positive' rationality, tenants of the fast-and-frugal-heuristics program consequently have a very different interpretation of 'normative' rationality. Instead of focusing on optimising processes, they advance the concept of *satisficing*: the idea that one choice may be 'good enough' in a given environment, thus making the concept of error/mistake meaningful when it leads to 'good' (satisfying) rather than 'bad' (non-optimal) choices.

Tenants of the fast-and-frugal-heuristics program then 'argue for an alternative non-axiomatic approach to normative analysis focused on veridical descriptions of decision process and a matching principle — between behavioral strategies and the environments in which they are used — referred to as *ecological rationality*' (Berg and Gigerenzer 2010, p. 133 — their emphasis). 'Ecological rationality' is a term they borrow from Smith (2003) that focuses on the question of which heuristics are adapted to which environments. The bottom line is that according to these authors, individuals *do* make 'good enough' decisions when they are constrained by what tenants of the heuristics-and-biases program judge to be a prejudice against individuals, e.g. limited attention and imperfect/incomplete information.

Since the fast-and-frugal-heuristics program (i) is grounded on a radically different approach of 'positive' behavioural economics than the heuristics-and-biases program, since (ii) the two programs are based on fundamental epistemic disagreements about

[8] For more about this debate, see Gigerenzer (1991), Kahneman and Tversky (1996) and Gigerenzer (1996).

the interpretation of probabilities (Bayesians *versus* Frequentists), and since (iii) they also disagree about the concept of normative rationality, it follows that it would have required to create another super-category in my taxonomy above that distinguishes all the normative approaches of the heuristics-and-biases program from the normative approach of the fast-and-frugal heuristics program. But because the fast-and-frugal-heuristics program is another most influential research program on its own, it would have required another full work to study its normative implications.

CHAPTER 1

From Prospect Theory to Behavioural Welfare Economics

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 chapter 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 chapter 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.

Acknowledgements. Submitted version to the *Journal of Economic Methodology* (July 2020). 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.

1.0 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 2015a; Lecouteux 2016; Moscati 2018 [Ch. 16]).[1]

The goal of this chapter 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*.[3] Second, PT is one

[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).

[3] As a matter of fact, two works explicitly consider PT to be the referent descriptive decision-making model for normative analysis (Bleichrodt, Pinto, and Wakker 2001; Pinto-Prades and Abellan-Perpiñan 2012). I come back to this point below.

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 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 chapter is organised as follows. Section 1.1 briefly introduces the components of PT by showing that they constitute an important matter of concern in BWE. Sections 1.2 and 1.3 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 1.2) and how the separation between descriptive and normative decision-making is useful to normative analysis (Section 1.3). Section 1.4 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 1.5 concludes.

1.1 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.[4]

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*

[4]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).

(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 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 1.2. 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.

1.1.1 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*.

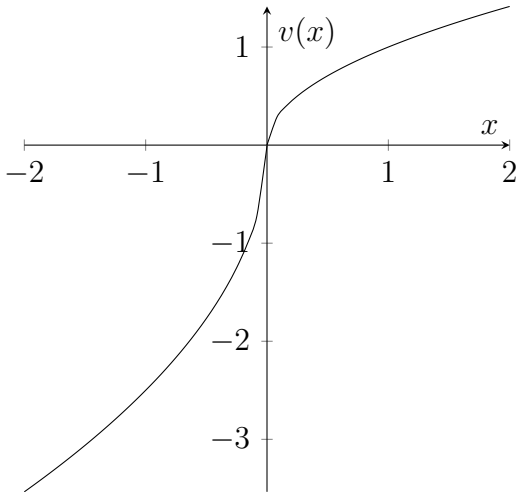
Consider first the reference point of PT. It is generally represented as the *status quo* and serves as the benchmark to distinguish gains from losses. It is assumed to be a neutral reference outcome, which is assigned the value of zero. One famous policy application which exploits this cognitive bias 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.[5]

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

[5]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.[6] [7]

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

[6] 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 $v(0)$.

[7] 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).

one 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.[8] 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.[9]

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).[10]

[8] 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.

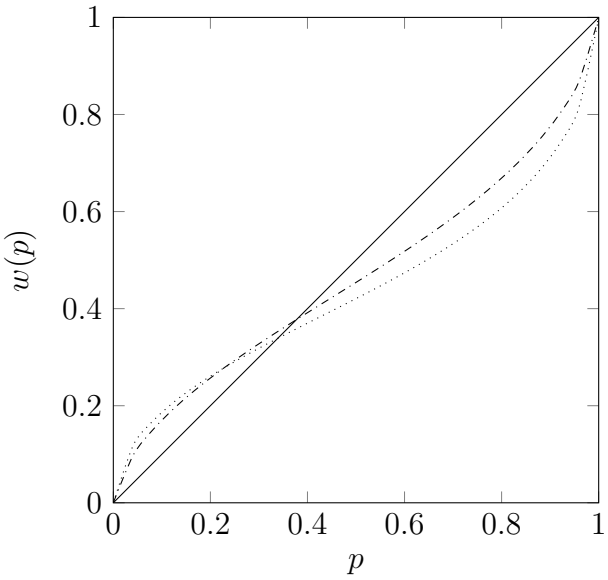
[9] Harrison and Ross (2017) provide an interesting criticism to the assumption that loss aversion is a prejudice against oneself. In their words, 'λ is a response operator, most naturally interpreted as reflecting a sentimental influence on behavior and cognition. To the extent that a person experiences 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.

[10] 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.[11]

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 mention

[11] 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.

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.

1.1.2 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.[12]

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 of view of behavioural

[12] 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).

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.

1.2 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.[13] 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.

[13] *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.

1.2.1 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.[14] 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

[14] 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).[15] 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*.

1.2.2 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)[16]

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'

[15] 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.

[16] 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.[17] 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.[18]

[17] 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).

[18] 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.[19]

1.3 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., 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)

[19] 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'.

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

1.3.1 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.

1.3.2 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).[20] The idea is if descriptive decision-making can apply to 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.

1.4 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 problem that are in fact equivalent when shown together should elicit the same preference,

[20] 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).

even when they are shown separately. As an illustration of violation of invariance, consider the following choice problem proposed by KT (1981, p. 454).[21]

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*).

[21] 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)[22]

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 think they have

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

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.

1.5 Conclusion

This chapter 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.[23] 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,

[23] 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 chapter.

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.

CHAPTER 2

Back to Aristotle? Explorations of Objective Happiness

Abstract

This chapter provides an analytical assessment of measuring experienced utility: a research program that leading expert Daniel Kahneman recently stated to have abandoned. My analysis follows four steps. First, I propose a literature review of twenty years of experienced utility measurement. Second, I consider several philosophical issues that Benthamian hedonism may be a problem for public policy. Third, I provide a philosophical discussion of all the axioms of experienced utility measurement by arguing that many of them suffer from important theoretical issues. Finally, I show that maximising individuals' moment utilities is based on a misconception of happiness that economists and policymakers have good reason to stay aware from. The bottom line is if economists and policymakers seek to improve their understanding on measuring objective happiness, Aristotle's eudaimonism may provide a more convincing account of objective happiness that palliates some of the issues of Bentham's hedonistic reductionism.

Acknowledgements. Version August 2020. I thank Jules Le Lay and David Lowing for fruitful discussions. I am also grateful to Jean-Sébastien Gharbi and Cyril Hédoin for careful reading. All mistakes remain my own.

2.0 Introduction

> *'Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do, as well as to determine what we shall do.'*
>
> Bentham (1780 [2007], p. 23)

> *'Now such a thing happiness, above all else, is held to be; for this we choose always for itself and never for the sake of something else, but honour, pleasure, reason, and every virtue we choose indeed for themselves (for if nothing resulted from them we should still choose each of them), but we choose them also for the sake of happiness, judging that through them we shall be happy. Happiness, on the other hand, no one chooses for the sake of these, nor, in general, for anything other than itself.'*
>
> Aristotle (-350 [2009], p. 10)

The essence of the present chapter can be resumed in the confrontation of these two different conceptions of happiness endorsed by Bentham, on the one hand, and Aristotle, on the other hand. Due to the many methodological and theoretical issues of measuring experiences of pain and pleasure, should we end up (once for all) with the research program of measuring experienced utility and focus instead on other meaningful factors of what constitutes 'objective' happiness? The main argument advanced in this chapter is that we should, for several reasons that lead me to show that experienced utility measurement is nothing but a dead end.

Measuring happiness through the hedonic interpretation Bentham (1780 [2007]) gave to it has been the subject matter of (approximatively) twenty years of research by Daniel Kahneman and his colleagues (henceforth 'Kahneman et al.'). The bulk of this research program was to concretise a utilitarian dream left unrealised by Edgeworth (1881):

> 'Let there be granted to the science of pleasure what is granted to the science of energy; to imagine an ideally perfect instrument, a psychophysical machine, continually registering the height of pleasure experienced by an individual. ... The continually indicated height is registered by photographic or other frictionless apparatus upon a uniformly moving vertical plane. Then the quantity of happiness between two epochs is represented by the area contained between the zero-line, perpendiculars thereto at the points corresponding to the epochs, and the curve traced by the index.' (p. 101)

However, despite all the efforts engaged by Kahneman et al. to concretise Edgeworth's 'hedonimeter' (to be reviewed in Section 2.1), experienced utility measurement is currently at its lowest profile. Perhaps the main reason for this is that such research program was abandoned by leading expert Daniel Kahneman himself. In a recent interview given to *Hareetz* (an Israeli online newspaper), the author explicitly declared that he does not believe anymore in the research program he undergone for twenty years. In Daniel Kahneman's own words,

> 'I gradually became convinced that people don't want to be happy ... They want to be satisfied with their life. People don't want to be happy the way I've defined the term — what I experience here and now. In my view, it's much more important for them to be satisfied, to

> experience life satisfaction, from the perspective of "what I remember", of the story they tell about their lives. I furthered the development of tools for understanding and advancing an asset that I think is important but most people aren't interested in.' (Daniel Kahneman — interviewed by Amir Mandel in 2018)[1]

The 'new' Daniel Kahneman (henceforth Kahneman-2018) recognises that he might have missed the point of what objective happiness truly is in giving *moment* utilities (what is experienced here and now) too much importance. Instead, what may fundamentally matter is not the *experienced* value of a decision but its *remembered* value: what individuals remember of their experiences. Yet maximising moment utilities for public policy was the whole point of Kahneman et al., particularly in Kahneman (1999, 2000). So what happened after twenty years of experienced utility measurement that Daniel Kahneman himself led to abandon the research program he undergone for all these years?

In this chapter, I do not provide a historical analysis of measuring experienced utility, which is a gap already fulfilled by Read (2007). Instead, my aim is to provide an analytical assessment about what is philosophically problematic with the whole theory of experienced utility measurement. Specifically, my aim is to (i) provide a literature review of the program of Kahneman et al. (from the beginning until its 'end'), (ii) consider several philosophical issues that Benthamian hedonism may be a problem for public policy, (iii) philosophically discuss all the axioms of experienced utility measurement, and ultimately (iv) explain that maximising individuals' moment utilities is based on a misconception of happiness that economists and policymakers have good reason to stay away from.

Because of all the methodological and theoretical issues of measuring experienced utility discussed throughout this chapter, it is argued that those issues give economists and policymakers reason for endorsing alternative measures of objective happiness *that are not directly related to the seek of pleasure*. Pleasure is undoubtedly a good thing. Yet perhaps the biggest mistake was to consider it, as Bentham (1780 [2007]) did, as what *constitutes* goodness (and more straightforwardly, the ultimate end of everything). The point is that after Kahneman-2018's own acknowledgement that maximising moment utility may not be what matters to the good life, a complete and up-to-date assessment of experienced utility measurement is needed. Do we have good reason to continue with Benthamian hedonistic reductionism as the ethical benchmark for public policy, and if not, what alternative direction can we propose to the program of objective happiness measurement?

Among the various conceptions of happiness, there is yet another one closely related to Kahneman-2018's statement in which pleasure is not what constitutes happiness but is either a *component*, a *process* or a *by-product* of it. Such conception is found in Aristotle's (-350 [2009]) *Nichomachian Ethics*, where happiness is also presented as the ultimate goal of life but where — in contrast with Benthamian hedonism — *happiness is not defined as the pursue of pleasure*. In Aristotle's (-350 [2009]) ethics, happiness is defined by the function of man: what she can successfully accomplish (p. 11). According to Aristotle, if happiness is a human quality then it needs to be located in the 'activity of soul which follows or implies reason' (p. 11). In his terms, happiness is about possessing the exercice of thinking about one's condition.

[1] Full journal article is available at https://www.haaretz.com/israel-news/.premium.MAGAZINE-why-nobel-prize-winner-daniel-kahneman-gave-up-on-happiness-1.6528513.

So how can the ethics of Aristotle enlighten us on the way objective happiness should be measured? The answer is by extending Kahneman-2018's reconsideration of objective happiness, which is to be broadened by integrating other components than what is experienced here and now. Those components may be the seek of human flourishing such as social relationships and overall life satisfaction (e.g. being in good health and having access to decent living conditions). It also requires to give importance to what individuals remember of their past experience, even if those experiences do not maximise pleasure. In this conception of objective happiness, pleasure may come as a by-product of exercising a 'virtuous' life *but pleasure is not what ultimately matters to the good life*.

The bottom line is if economists and policymakers are willing to find a proper measure of objective happiness, they have perhaps good interest in looking for alternative normative approaches that do not make of pleasure maximisation their ultimate goal. A normative approach which is already well aligned with Aristotle's conception of the good is actually the capability approach (Sen 1985; Nussbaum and Sen 1993; Nussbaum 2000). Such normative approach is given brief consideration in conclusion.

The rest of the chapter is organised as follows. Section 2.1 provides a literature review of measuring experienced utility by the program of Kahneman et al.[2] Section 2.2 discusses several reasons that Benthamian hedonism may be problematic for public policy. Section 2.3 philosophically discusses each of the ten axioms of the theory of experienced utility, which permits its measurement. Section 2.4 reconsiders the content of experienced utility measurement, i.e. what matters may not be *moment* utility but *remembered* utility. I then conclude in Section 2.5 with a brief appraisal of Aristotle's eudaimonism, which may provide a richer conception of objective happiness measurement than Bentham's hedonistic reductionism.

2.1 Measuring Experienced Utility: A Literature Review

The discrepancy between 'decision utility' and 'experienced utility' was initially suggested by March (1978), who made the case that decision value and experience value typically do not converge for ordinary decision makers. *Decision* utility is the weight of an outcome in a decision, as in any model of decision-making. *Experienced* utility is the hedonic quality as in Bentham's (1780 [2007]) usage. That is, it is the experience in term of happiness, which is not necessarily related to one's observed choice. Since decision utility is inferred from observed choices, and since observed choices are sometimes subject to cognitive biases, the idea is that individuals may not always choose the outcome that makes them better off.

The conceptual appeal of experienced utility is to consider it as a more reliable proxy of well-being than decision utility and to use such normative criterion for public policy. The main advantage of the experienced utility criterion is that it is independent of the choices

[2]Awareness should be given that my aim is not to provide a *general* literature review of happiness measurement. The reader can find enough material in Frey and Stutzer's (2002) state of the art, in Layard's (2011) enthusiastic defence of making happiness the central criterion for public policy, and in Read's (2007) insightful history of the concept of experienced utility 'from Jeremy Bentham to Daniel Kahneman'. See also Angner (2013) for an assessment of the possibility of measuring a mental state account of well-being defined in terms of happiness.

individuals make, and hence can be used to evaluate which choices increase well-being and which choices decrease it. The separation between decision utility and experienced utility was already a matter of discussion in Kahneman and Tversky (1984, pp. 349-350) and became the central interest of the research program of Kahneman et al. concerned in evaluating the process of individuals when they endure experiences of *pain* and *pleasure*.

The main interest of this research program was to understand the connection and gap between what individuals *experience* in real time — i.e. the way they actually lead their life — and what they *remember* of those experiences — i.e. the narrative they represent themselves about the way they lead their life. Tenants of the experienced utility criterion for normative assessments argue that policy recommendation can be based on the evaluation of *total utility*: the collection of utility profiles which follows certain normative rules (that are explained below). According to the experienced utility criterion, a policy is judged to be better than another if it maximises the level of total utility. Its central ethical rule can then be formulated in the following premise.

Ethical premise. *An individual's state of affairs is better than another if it has more level of total utility than another*. Formally, let $x = (x_1, ..., x_n) \subseteq X$ be a realisable set of an individual's states of affairs (e.g. a consumption bundle, health states, sips of tea, etc.) and X be the set of outcomes. I denote by $i = \{0, ...n\}$ the index of time for each element of the vector x. For example, x_1 is one sip of tea at time 1, x_2 another sip of tea at time 2, and so on. $W(x)$ is an individual welfare function of the form,

$$W(x) = \int_0^n u(x_i)dx$$

where $u(x_i)$ is the individual's utility profile of x at time i and $\int$ the integral of all utility profiles, which simply allows to have the total utility of the individual.[3] The experienced utility criterion is satisfied under the condition that,

$$W(x) > W(x') \implies x \succeq x'$$

Since 1990, the development of the experienced utility criterion can be resumed in three main lines: (i) theoretical distinctions between several kinds of utilities (*decision utility*, *experienced utility*, *predicted utility*, *moment utility*, *remembered utility* and *total utility*), (ii) accumulation of empirical evidence about the way individuals perceive and remember experiences of pain and pleasure, and (iii) methodological improvements of measuring the aggregation of moment utilities. The individual contributions are the following.[4]

2.1.1 Individual Contributions

Kahneman and Snell (1990) provide empirical evidence of generally poor performance in the task of predicting utility. They introduce the concept of *predicted utility*, defined as the belief one has about future experienced utility. The authors argue that if individuals fail to anticipate the effect of the outcome of a choice on their future preferences and

[3] Although it is not absurd to use a sum, the integral better captures the summation of utility profiles because such summation graphically represents an 'area' of pleasure (or pain) if we consider time to be a continuous variable.

[4] If the reader is already well aware of this literature, she may safely skip to Section 2.2.

hedonic states, decision utility and experienced utility should be treated differently.

Kahneman and Varey (1991) discuss the status of choice as the sole measure of utility and argue that deriving utilities from preferences is questionable. They suggest a distinction of experienced utility in three separate factors: the experience *as it happens*, the experience of *remembering it* and the experience of *anticipating it*.[5] They consider two important determinants of experienced utility: processes of *adaptation* and processes of *comparative judgement*. According to the authors, one important implication of adaptation is that it allows to make interpersonal comparisons of utilities when two individuals who are fully adapted to different levels of stimulation can be said to be matched in their absence of response to their states.

Also, Kahneman and Varey (1991) advance that if their responses to stimuli differ in the same direction from their respective adaptation levels, those can be matched in signs, if not in magnitude (p. 138). According to the authors, comparative judgements are more salient when individuals identify a reference person or group similar to themselves. They then advance that comparative judgements also allow for interpersonal comparisons of utilities when a group of similar individuals (e.g. poor or riches) see a change in their initial endowment. The authors argue that the history of prior experiences and the context to which the relevant object, state, or event is to be compared are likely to affect experienced utility, and that any treatment of interpersonal comparisons of utilities should give importance to these two factors.

Kahneman and Snell (1992) address the following question: 'do decision makers accurately predict their future hedonic experiences?' The empirical answer they provide is negative. In a series of recurrent experiments corresponding to eight consecutive days, subjects were asked to consume their favorite ice cream while listening to the same piece of rock music. After each episode, subjects had to rate how much they liked the ice cream and the music. At the end of the first session, they had to predict the ratings they would make on the following day and on the final day of experiment. The authors find a correlation between actual and predicted changes in liking close to zero. In sum, subjects poorly predicted their future hedonic experiences. Although Kahneman and Snell (1992) give full awareness that their results are insufficient to claim that people have trouble in predicting their future tastes, they argue to be sufficient in order to indicate failures in such tasks.

Varey and Kahneman (1992) address another important question: 'do people correctly incorporate their hedonic beliefs into their decisions?' The authors propose 'utility integration' as a normative standard, which takes '*the sum of the hedonic values associated with the separate moments as the measure of the experienced utility (or disutility) of the series*' (p. 170 — their emphasis).

In their view, utility integration should satisfy three conditions: *monotonicity* (or dominance), the rule according to which adding pain to a series should strictly increase global disutility; *non-discrimination*: two moment-pain experiences of equivalent magnitudes should be considered equally unpleasant contributions to the series; *additivity*:

[5]This actually corresponds to Jevon's (1905) enumeration of three distinct ways in which pleasure or pain are caused. See also Loewenstein (1987) who study the value that individuals attribute on waiting periods in which to enjoy or to suffer anticipation of future hedonic events.

the difference between the global disutility of an unpleasant experience at i and the one of an unpleasant experience at $i+1$ is simply the disutility of the extra unpleasant experience. In their experiments, subjects have to endure painful experiences such as carrying a suitcase, sitting in a vibrating room or standing in an uncomfortable position. Their experiments show that most of the subjects violate utility integration. One of their important findings is that adding pain to a series can produce a lower global evaluation, which is not in accord with monotonicity.

Kahneman et al. (1993) show that actions that are based on memories of experiences which have systematic biases relative to contemporaneous evaluations of experiences may strongly support an interpretation of mistake. In their experiment, subjects have their hand submerged into cold water. There are two settings: one shorter duration (60 seconds) at 14°C and one in which an extra duration time is added (+30 seconds), where the temperature is slightly increased to 15°C. The empirical results showed (again) that subjects violated temporal monotonicity — the rule according to which adding moments of pain to the end of an episode makes it worse, and that adding moments of pleasure makes it better.[6]

Fredrickson and Kahneman (1993) show identical results with snapshots, where people were exposed to sixteen short plotless film clips, half pleasant (e.g. views of coral reef) and half unpleasant (e.g. an amputation). Schreiber and Kahneman (2000) provide further empirical support for such result with aversive sounds of varying loudness and duration, so as Redelmeier, Katz, and Kahneman (2003) who in a randomised trial assign to half of the patients an added short interval to the end of their colonoscopy.

Kahneman (1994) summarises three important points known from empirical research: (i) people are myopic in their decisions, (ii) they may incorrectly predict their future tastes and (iii) they make erroneous choices by fallible memory and incorrect evaluation of past experiences. Due to these observations, the author argues for an enriched definition of rationality with what he calls the 'substantive' criterion of experienced utility: a criterion that evaluates the outcomes of decisions *independently from (or external to) the system of preferences*. This constitutes an important departure from standard welfare economics, which is based on the satisfaction of individuals' *preferences* in order to evaluate their states of affairs.

Kahneman (1994) then introduces two empirical generalisations known as (i) the *peak-end rule*: global evaluations of experiences are accurately predicted by the mean between the most unpleasant feeling in the episode and the one recorded at the end of the episode; (ii) *duration neglect*: the duration of an unpleasant episode has no significant effect to retrospective evaluations of experiences. These two conclusions particularly originate from an experiment of Redelmeier and Kahneman (1996) about the intensity of pain experienced by patients undergoing colonoscopy.

[6]Note however that Varey and Kahneman (1992) only define monotonicity in terms of pain, while the definition of monotonicity in terms of pain *and* pleasure is taken for granted in Kahneman et al. (1993). However, one may argue that monotonicity cannot account for *both* painful and pleasurable experiences because their remembered perception can be interpreted differently by the subjects. Empirical evidence of the peak-end rule (see below) in terms of *pleasure* (and not pain) is actually scarce (Do, Rupert, and Wolford 2008), if not non-existent (Kemp, Burt, and Furneaux 2008; Mah and Bernstein 2019).

Kahneman, Wakker, and Sarin (1997) in their seminal 'back-to-Bentham' approach propose a formal normative theory of what they call the *total experienced utility of temporally extended outcomes*: a sequence of life experiences that can include anything related to the sensation of pleasure and pain. The authors aim at measuring what they call 'temporally extended outcomes' (TEOs) with the normative concept of 'total utility': an aggregation of temporal profiles of utility which is experienced instantly by individuals.

They first provide empirical evidence that the system in which normal individuals form and store evaluations of situations is not designed to optimise experienced utility. Then, they propose a normative theory from the concept of 'total utility'. The authors aim to specify 'the conditions under which the total utility of an extended outcome is the temporal integral of some transformation of instant utility' (p. 388). They suggest that a policymaker could eventually maximise the sum of the total utility of each individual into an objective function.

Kahneman (1999) explores the concept of objective happiness, an attempt to specify what an external observer would need to know in order to determine how happy an individual is at a given period, and the rules for using that knowledge. According to Kahneman (1999), the highest level of evaluating well-being is grounded on information about *instant* (or moment) utility. The author argues for a 'bottom-up' construction of individuals' global evaluations of well-being by distinguishing two notions of happiness: *subjective happiness*, based on self-stated 'how happy are you' reports and *objective happiness*, derived from a record of instant utility over the relevant period.

The author states that remembered utilities and total utility of episodes differ just as subjective and objective happiness: the former gives an approximate evaluation of one's well-being, while the latter gives a more precise valuation of happiness. Although objective happiness is naturally determined by subjective self-reports, the idea is that the aggregation of instant utility is governed by a logical rule that is *external* to the subject, i.e. a rule stated by the social planner, just like in Kahneman, Wakker, and Sarin (1997).

From Kahneman's (1999) viewpoint, only *objective* happiness is normatively relevant. This has major implications for public policy. As the author claims, 'policies that improve the frequencies of good experiences and reduce the incidences of bad ones should be pursued even if people do not describe themselves as happier or more satisfied' (p. 15). He explicitly argues that the goal of policy should be to increase measures of *objective* happiness, not measures of satisfaction or *subjective* happiness.

Kahneman (2000) presents an overview of the experienced utility criterion and of the relation between the pleasure and pain of moments and the utility of more extended episodes. The author argues that experienced utility is better measured by moment-based methods — that assess the experience of the *present* — rather than by the memory-based approach — which takes the subject's retrospective evaluation of *past* episodes (remembered utility) as valid data. He then develops his concept of 'objective happiness' that he already introduced in Kahneman (1999).

The author argues that the general distinction between decision utility and experienced utility has major implication to normative assessments in public policy. Taking

typical questions of cost-benefit analysis — e.g. 'Does the presence of trees in a city street affect the mood of pedestrians?', 'What are the well-being consequences of inflation, unemployment, or unreliable health insurance?' (p. 204) — Kahneman (2000) argues that in addition to standard methods of willingness to pay/willingess to accept and elicitation of public opinion, there is a substantial interest in measuring the experienced utility associated with public goods.

Kahneman et al. (2004) introduce the Day Reconstruction Method (DRM): an alternative measure of subjective-well being that they argue to palliate the issues of previous sampling methods of experienced utility. Before the DRM was the privileged measure of subjective well-being, measuring experienced utility was possible with Experience Sampling Methodology (ESM) (Larson and Csikszentmihalyi 1983). Respondents in ESM studies are asked, with the help of a palmtop computer they carry along the day and which beeps at random times, to record where they are, what they are doing, and how they feel several times throughout the day.

The aim is to collect the most accurate data possible by targeting multiple and immediate reports from people in their typical environments. According to Kahneman et al. (2004), the disadvantage of the ESM is that experience sampling is expensive, involves high levels of participant burden, and provides little information about uncommon or brief events, which are rarely sampled. Instead, they argue that the advantages of the DRM are that it imposes less respondent burden, does not disrupt normal activities, and provides an assessment of contiguous episodes over a full day, rather than a sampling of moments (p. 1777).

The DRM consists in asking respondents to first revive memories of the previous day by constructing a diary consisting of a sequence of episodes: 'Think of your day as a continuous series of scenes or episodes in a film'; 'Give each episode a brief name that will help you remember it (e.g. commuting to work, at lunch with your colleague, etc.)'. Then, respondents are asked to describe each episode by answering questions about the situation and about the feelings that they experienced, as in experience sampling.

Kahneman and Sugden (2005) aim to explore the implications of basing economic policy evaluation on experienced utility. The authors discuss the problem of contingent valuation when based on the standard method of willingness to pay and willingness to accept. One central concern they discuss is to evaluate states of affairs on stated preferences, while stated preferences might be subject to cognitive biases. For example, when it is asked to individuals to think about what it would be like to be in some continuing state (e.g. living in California or being paraplegic), what they actually think about is what it *would be like to move to* that state, not what it *is like to be* in that state.

This heuristic refers to the 'transition heuristic': the failure of taking into account adaptation. But if people do not anticipate adaptation, the authors argue that responses to stated preference questions may reflect systematically bias forecasts of experienced utility. A second bias they mention is the 'focusing illusion': 'nothing in life is as important as you think it is when you're thinking about it' (Schkade and Kahneman 1998). For example, when we are thinking about a paraplegic person, we are thinking about that person *thinking she is a paraplegic*. But empirical evidence showed no significative decrease in

subjective well-being for paraplegic individuals, simply because they tend to forget being paraplegic in the long run (Brickman, Coates, and Janoff-Bulman 1978).

The point is that people may overestimate the effect of a particular state of affairs because they attribute too much attention to this state of affair. As a consequence, the authors argue that this phenomenon may ultimately bias subjective reports. Like Kahneman et al. (2004), Kahneman and Sugden (2005) defend the DRM (well-being measured in terms of moment-based utilities) as a better alternative than anticipated utility and overall satisfaction measures. They argue the DRM to be used to estimate the effects on happiness of many kinds of goods that are currently subject of contingent valuation, such as landscapes, recreation sites and states of health.[7]

Kahneman and Krueger (2006) discuss how individuals' responses to subjective well-being questions vary with their circumstances and other factors. Like in the previous works, they argue for a necessary distinction between different conceptions of utility rather than a single one. The novelty proposed in this paper is the 'U-index' defined as 'a misery index of sorts, which is the proportion of time that people spend in an unpleasant state' (p. 4). According to the authors, the U-index avoids the difficulty of giving a cardinal representation of utilities for making interpersonal comparison because it actually provides an *ordinal measure at the level of feelings*. The U-index (for 'unpleasant' or 'undesirable') is constructed as follows.

The authors first classify an episode as unpleasant if the most intense feeling reported for that episode is a negative one. In other words, if the highest rating on any of the negative affect dimensions is strictly greater than the maximum of rating of the positive affect dimensions, then such episode is a negative one. In doing so, it does not matter whether two individuals who are differently sensitive to emotional states use, say, the 2 to 4 portion of the 0 to 6 intensity scale of unpleasant states (individual 1) and the full range of the scale (individual 2). As the authors put it, as long as both individuals 'employ the same personal interpretation of scales to report the intensity of positive and negative emotions, the determination of which emotion was strongest is unaffected (ignoring ties)' (p. 19).

According to the authors, this method has three main advantages. First, it only requires one salient negative emotion for an episode to be unpleasant. Since individuals mostly endure a positive predominant emotional state in an episode, one 'extreme' negative emotion provides a significative and contrasting occurrence. Second, selecting a negative feeling as more intense to a positive feeling is likely to be a deliberate choice because negative feelings of this sort are relatively rare (at least for individuals living in rich and developed countries). Third, the correlation of the intensity among various positive emotions across episodes (e.g. 'being happy' and 'enjoying oneself') is higher than the correlation among negative emotions (e.g. 'being depressed' and 'feeling angry'). According to the authors, this also provides significant and contrasting salience on how

[7]The two authors however conclude with diverging opinions regarding the future of experienced utility for normative assessments. While Kahneman is enthusiastic, Sugden is more skeptical, arguing that the aim of public policy is rather to promote institutional arrangement so that individuals can purchase goods and services that they are willing to pay for, even if preferences fail to meet conventional consistency conditions, and even if preference-satisfaction conflicts with well-being. For a critical review of Sugden's (2018a) normative approach, see Mitrouchev (2019).

the subject experiences the entire episode.

2.1.2 The End of Experienced Utility Measurement?

Since the end of the 2000s, the experienced utility criterion made no more significant empirical and theoretical improvements. Other contributions either provide literature reviews which document errors of hedonic forecasting (Kahneman and Thaler 2006; Dolan and Kahneman 2008), discuss its practical issues (Loewenstein and Ubel 2008), provide empirical tests of the peak-end rule (Do, Rupert, and Wolford 2008; Kemp, Burt, and Furneaux 2008; Mah and Bernstein 2019), provide empirical tests of the fundamental distinction between decision utility and experienced utility (Carter and McBride 2013; Akay, Bargain, and Jara 2017), are made for the public reader (Kahneman 2011 [Part V]), or focus on particular epistemic issues about measuring health states (Hausman 2015; Oliver 2017).

With what has been said so far, we have enough material to discuss the methodological and theoretical issues of the experienced utility criterion. The next section underlines several problematic principles of grounding public policy on Benthamian hedonism. Section 2.3 discusses the theory of experienced utility measurement. Section 2.4 then reconsiders the content of experienced utility measurement (*remembered* utility instead of *moment* utility).

2.2 Why Hedonism May Be a Problem for Public Policy

Evaluating individuals' level of happiness with the experienced utility criterion invokes a conception of the good life, but quite a peculiar one: it is a good thing to maximise individuals' *moment* (or instant) utilities. The underlying theory of well-being on which this criterion is grounded is a particular form of hedonism considered to be equivalent with Benthamian utilitarianism. But this ethical theory may appear to be 'narrow' in the sense that it excludes a lot of human considerations about what makes the good life.

First, Benthamian utilitarianism is in fact only a particular form of hedonism, according to which the virtue of life is to *maximise pleasure* and to *minimise pain*. But the roots of hedonism can be traced back to the Cyreanic school, where no pleasure/pain calculus was at the time a matter of concern. Also, although Epicurus (B.C. [1994]) — known as being a forerunner of hedonism — proposed several prudential rules for reaching *ataraxia* (the absence of mind troubles), he certainly did not refer to a rational calculus of pain and pleasure in the same way than Bentham (1780 [2007]).

Second, when other values such as freedom, fairness, compassion, equality and rights are involved, it is well acknowledged that hedonic value is of no use to assess individuals' states of affairs. For example, Sen (1991, p. 25) argues that it is important to take into account in the concept of preference-satisfaction that individuals have 'the freedom to lead the life [they] would choose to lead' by bringing counterfuctual choices (what they *would have chosen*) into the evaluation.

Third, in an empirical study Smith et al. (2006) found that colostomy patients reported similar levels of happiness to people who did not have colostomies. However, colostomy

patients also expressed a willingness to give up 15% of their remaining life span if it could be lived without colostomy. This indicates that those patients placed a high value on having their former health restored, which indicates in return that important human values such as being in good health are neglected within the experienced utility approach.

Fourth, we can refer to Nozick's (1974) 'pleasure machine' thought experiment, which consists in asking whether we would prefer to be connected to a machine that would maximise our experienced utility rather than living the real life. Nozick (1974) provides three arguments why it is not desirable to do so. First, we want to *do* certain things, not just have the experience of doing them. Second, (in relation to the first point), this is because we want to be a certain kind of person and not 'an indeterminate blob floating in a tank' (p. 43). Third, plugging into an experience machine limits us to man-made reality, where there is no contact with a 'deep reality'.

Although intuitively appealing, the overall criticism that experienced utility is 'too narrow' to capture what makes the good life is actually the easiest and perhaps the less relevant one. In fact, it is important to note that tenants of the experienced utility criterion for normative assessments fully acknowledge this normative criterion to be only relevant to *particular* circumstances, and that Benthamian hedonism — the ethical theory on which this normative criterion is grounded — should not be taken at face value.[8]

Their argument is that hedonism is a *component* of what constitutes the good life. In this sense, it is not in conflict with other values such as freedom or fairness. They claim that the evaluation of hedonic states is surely not adapted to every circumstances, but its usefulness is certainly *not empty* regarding some situations where 'a separate value judgment designates experienced utility as a relevant criterion for evaluating outcomes' (Kahneman, Wakker, and Sarin 1997, p. 377). The intuition behind this argument implicitly holds under two conditions, which, if satisfied, make experienced utility a good guide to well-being in *some* circumstances.

- *Condition 1*. There exist cases in which the evaluation of states of affairs refer to hedonic states (such as the selection of an ice cream flavour) and not other things.
- *Condition 2*. Those cases are intuitively known (at least approximatively).

I shall seriously consider these two conditions in turn.

2.2.1 Assessment of Condition 1

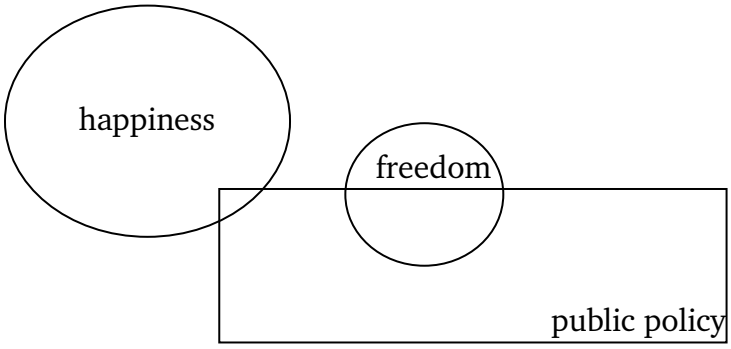
There is potentially a consequent number of public policies which can be concerned with the promotion of happiness. But the issue may not be that hedonic maximisation does not apply for many cases in life. Instead, if we think that public policy is not concerned about happiness defined in terms of intensities of pleasure but in other terms (e.g. overall satisfaction of one's life or democratic participation), then the experienced utility criterion may be a restrictive normative criterion for public policy.

[8] See Varey and Kahneman (1992, p. 169), Kahneman (1994, p. 21), Kahneman, Wakker and Sarin (1997, p. 377) and Kahneman and Sugden (2005, p. 176).

My present argument basically says that the scope of what matters to individuals *in life* and what public policy can do about making individuals' life better — i.e. the scope of *public policy* — are not necessarily equivalent. In other words, even if the set of what matters to individuals in life largely includes happiness defined in terms of maximisation of pleasure and minimisation of pain, it does not mean that public policy is (or can be) particularly concerned with this social dimension.[9]

An illustration is given in the following diagram with the hypothetical representation of three sets: *happiness* and *freedom* (belonging to the superset of values) and *public policy* (belonging to the superset of the practical dimension of life, as in Aristotelian terms).

Figure 2.1: Hypothetical representation of 'happiness', 'freedom' and 'public policy' sets



Imagine, for the sake of argument, that we live in a world where happiness matters more than freedom (i.e. where the set of happiness is bigger than the set of freedom).[10] If the intersection between happiness and public policy is smaller than the one between freedom and public policy — i.e. the domains of life in which the policymaker can actually do something about individuals' states of affairs — we might have reason to doubt about the extensive scope of the experienced utility criterion for public policy.

As Kahneman and Sugden (2005) put it,

> 'even if one accepts experienced utility as a measure of well-being, one may ask whether it is a government's job to create well-being for its citizens.' (p. 177)

How big the intersection between 'happiness' and 'public policy' is can be either determined by empirical evidence (how many public policies actually aim at maximising individuals' experienced utility?) or, I think more interestingly, by a philosophical assessment (what is public policy merely about?). In fact, whether a particular domain of policy is to be categorised as either more 'happiness-relevant' or more 'freedom-relevant'

[9] Some authors usually take for granted the utilitarian ethical judgement that the goal of any public policy ought to maximise individuals' well-being (Loewenstein and Ubel 2008, p. 1804; Dalton and Ghosal 2011, p. 565). But compared to any other approach in political philosophy, this utilitarian view is far from being self-evident.

[10] For simplicity, I assume in the above diagram that the intersection between happiness and freedom is empty. But happiness has surely something to do with freedom, depending on what we actually mean by 'happiness'. For an empirical test of the correlation between happiness and freedom in 38 mainly developed nations at the beginning of the 1990s, see Veenhoven (2000).

is not that clear.

Consider for example the 401(k) default option policy, which aims at increasing the number of employees' enrolment so that employees increase the total amount of saving for their retirement (Madrian and Shea 2001; Thaler and Benartzi 2004; Bernheim, Fradkin, and Popov 2015). Does enhancing employees' saving yield to more *happiness* or more *freedom*, considering that they will enjoy a larger amount of money when they will be retired? How about the time-selves employees who save at each period? Is the policy more relevant regarding the anticipated happiness it produces or regarding the anticipated freedom it produces?

The experiences of Kahneman et al. reviewed in Section 2.1 are designed in situations where patients undergo colonoscopies, sit in a vibrating room, stand in an uncomfortable position, hold their hand into cold water, eat an ice-cream while listening to a piece of rock music, etc. Whether these experiments have something in common with the range of public policy so that it can be claimed that they are relevant to the latter is up to debate.

In fact, public policy seems rather to be about promoting *indirect* factors of happiness such as giving people more freedom to participate in the democratic life, more opportunity to engage in the free exchange of goods, services, and labor, and more freedom in one's private life (e.g. to practice one's religion, to travel, or to get married) (Frey and Stutzer 2002, p. 423). In short, it is important to question whether the experienced utility criterion accommodates well with the range of what public policy is actually concerned with.

2.2.2 Assessment of Condition 2

In response to the statement of Kahneman, Wakker, and Sarin (1997) that 'a separate value judgment designates experienced utility as a relevant criterion for evaluating outcomes' (p. 377), one may object, as Fumagalli (2013), that 'the issue is precisely when this is the case, and by means of what criteria we are supposed to identify these situations' (p. 341).

Although this seems a relevant point, Fumagalli's (2013) scepticism would merely have to apply for every other normative criterion, i.e. not only a normative criterion which is concerned with happiness but also one concerned with other values such as freedom or fairness. Otherwise, it would be presumptuous to argue that we have a better intuitive perception of e.g. freedom and fairness rather than happiness. Thus instead of asking the question of when the specific circumstances of using one normative criterion over another are met, the issue seems to be whether those circumstances are scarce or abundant (*cf. condition 1*). Except arguing that decisions are never controlled by hedonic predictions — which is an extreme view that perhaps only few theories of ethics would hold — Fumagalli's (2013) objection is not to be taken at face value since every ethical representation of what makes the good life (at least formulated into a normative criterion) is necessarily *partial*.

That is, except having a normative criterion that can entail many different values of what matters to individuals (such criterion would have to be based on an ethical theory

that entails those different values), it seems that partial representations of what makes individuals better off are perfectly fine. I let the reader be the own judge of the two conditions here discussed, which (to me) are not the most concerning points one can make about the experienced utility criterion. More concerning, I believe, are the following theoretical issues.

2.3 Axioms of Utility Integration Are Debatable[11]

The construction of the temporal integral of moment utilities relies on six assumptions about subjects' ratings of instant utility (Kahneman 2000) and on four additional assumptions about a social planner who has a knowledge of the scale (Kahneman, Wakker, and Sarin 1997). Axioms 1, 2, 3 and 4 impose requirements on the measure of moment utility. They are epistemic judgements made for the practical usefulness of measuring total utility. Axioms 5 and 6 are normative rules which specify how total utility is constructed from moment utilities. They are ethical judgements made for summing moment utilities into total utility (or into an individual welfare function). Axioms 7, 8, 9 and 10 are technical assumptions about the transformation of utility profiles. In the present section, I discuss each axiom in turn.

2.3.1 AXIOM 1 (Inclusiveness)

Ratings must contain all the relevant information required for its temporal integral to be a plausible measure of the total utility.

This axiom merely consists in bounding the welfare-relevant domain. The informational basis of the experienced utility criterion is moment utility (what is experienced here and now). Note that moment utility also includes the affective consequences of prior events (e.g. adaptation, fatigue) and future events (e.g. fear, hope). This is an important characteristic for understanding (and perhaps also criticising) Axiom 5 (separability) below. Disputing the informational basis of the welfare-relevant domain (here moment utility) would lead us back to Section 2.2, so I move on.

2.3.2 AXIOM 2 (Ordinal Measurement across Situations)

The measurement of positive and negative deviations from zero is ordinal.

By definition, moment utility is the valence (good or bad) and the intensity (mild to extreme) of current affective or hedonic experience. This axiom basically says that the valence and intensity of a stubbed toe can be compared with the ones of a humiliating rebuke. For example, a pain rating of 7 in one situation (e.g. a stubbed toe) is considered of being worse than a pain rating of 6 in another situation (e.g. a humiliating rebuke), but the interval between 6 and 7 need not be psychologically equivalent with the interval between 3 and 2, although they must be measured on a common scale.

[11] A useful glossary of the experienced utility criterion is provided in Appendix 2.A, which resumes the technical concepts involved in its theoretical construction. I can recommend the reader to have a look at Appendix 2.A before reading this section, which discusses the theory of the experienced utility criterion.

This axiom can be disturbing to some because it requires to accept that different psychological perceptions (e.g. a stubbed toe and a humiliating rebuke) are categorised under a similar hedonic feeling. That is, one first needs to consider that a humiliating rebuke can be categorised as a (negative) hedonic feeling at all. In fact, whether both psychological phenomena of hedonic feeling (e.g. physical pain) and emotional feeling (e.g. emotional pain) are assumed to be commensurable is unclear in Kahneman, Wakker, and Sarin (1997).

The authors mention the affective experience of plotless film clips of Fredrickson and Kahneman (1993) to support the observation that individuals violate monotonicity (the rule according to which adding a moment of pain should reduce individuals' total utility). However, they make it quite explicit that their normative theory only applies to hedonic states that are naturally interpreted in terms of *physical* pleasure and pain, e.g. enjoying the taste of an ice cream or suffering a colonoscopy procedure. Kahneman (2000) instead considers both psychological phenomena of hedonic *and* affective experiences to be commensurable, which is, after all, a natural extension of the experienced utility criterion since it ultimately aims at being applied to public policy.

Kahneman (2000) advances that 'reporting the sign and intensity of current hedonic and affective experience is not essentially different from the standard psychophysical tasks of reporting color or smell' (p. 195). Here the author refers to his discussion in Kahneman (1999) about the large body of empirical studies in psychology on how human sensory experience works. Discussing this literature is outside the scope of the present chapter. What is relevant to my concern is the analogy Kahneman (1999) makes between the human sensory system and hedonic states.

The author acknowledges an important difficulty: 'it is more difficult — but not impossible — to compare the loudness of sounds that differ in pitch and in timbre than to compare sounds that share these attributes'. He then argues that 'the question of whether people can compare physical and emotional pain, or the trills of food and music is ultimately empirical' (p. 197). As I have no doubt that empirical studies can enlighten us on many psychological phenomena that are not fully understandable to humankind, and that the human sensory system is surely linked to our emotional responses, Kahneman's question seems also quite philosophical. Can empirical evidence actually tell us whether the sensation of eating chocolate while being sad provides meaningful comparison between the first and the second feeling?

I suspect many scholars to disagree with Kahneman's oversimplification, according to which almost every psychological perception can fit into a 'good-bad' scale (to be discussed below). Further empirical and philosophical assessments could perhaps enlighten us on this point, which is a complex debate very much linked with the assumption of interpersonal comparisons of utilities (to be discussed below).

2.3.3 AXIOM 3 (Distinctive Neutral Point)

The scale has a stable and distinctive zero point ('neither good nor bad', 'neither pleasant nor unpleasant'), which permits comparisons across outcomes and individuals.

This axiom is very familiar with the notion of a reference point, like in any reference-dependent model of decision-making, e.g. prospect theory (Tversky and Kahneman 1992). In prospect theory, the reference point generally represents the *status quo* and serves as the benchmark to distinguish gains from losses. Following the same logic, the neutral point of the normative theory of Kahneman, Wakker, and Sarin (1997) serves as the benchmark to distinguish positive from negative feelings.[12]

Some of course may argue that the existence of such neutral point is a strong assumption, since when asking an individual to evaluate her happiness, we in fact ask her to imagine an abstract state in which she evaluates her current mood from a zero point. But perhaps the main problem is that this neutral point may be changing. Typically, how can an individual who adapt to her life circumstances can evaluate a similar perception of pain and pleasure than before? Surely if she becomes rich to the point that she does not derive the same level of pleasure in eating tuna than before (because she can now afford caviar), it is hard to imagine that her hedonic level would not change according to her new circumstances.

Kahneman (1999, pp. 11-15) extensively discusses this point, recognising it to be difficult but not impossible to overcome. The main argument of Kahneman, Wakker, and Sarin (1997) is that 'the stimulus that gives rise to a neutral experience may be different in different contexts, but the neutral experience itself is constant' (p. 380). Thus, if we succeed to isolate individuals in an experimental setting for a short enough interval of time so that they do not have time to durably adapt, the evaluation of pain in relation to a neutral point may not be that problematic (e.g. as in the colonoscopy experiment of Redelmeier and Kahneman (1996)). The issue is that experienced utility measurement would then be quite restrictive if it cannot be applied to situations where these conditions are not satisfied.

Another point related to what has been said previously is that the 'bottom-up' construction of objective happiness (Kahneman 1999) requires that each moment can uniquely be characterised by a value on the 'good-bad' dimension. This technically requires the following two assumptions.

Assumption 1. The brain continuously constructs an affective or hedonic commentary on the current states of affairs — an assumption judged to be fairly supported by empirical evidence according to Kahneman (1999, 2000). In other words, any moment of time can be characterised by a particular value of the 'good-bad' dimension (positive, neutral or negative) but an evaluation cannot be both good and bad at the same time nor major manifestations of the 'good-bad' dimension can be dissociated.

Assumption 2. It follows from *assumption 1* that a commentary is adequately summarised by a single value, judged to be a 'tolerable oversimplificaiton' by Kahneman (1999, p. 7).

[12]Abstraction is made of any additional content associated with reference point, such as loss aversion. See Kahneman (1999, p. 18) for a discussion and Carter and McBride (2013) for an empirical test of whether the value function of prospect theory is of similar S-shape than the experienced utility function of Kahneman, Wakker, and Sarin (1997). The authors found mixed evidence for loss aversion in experienced utility.

Since *assumption 2* cannot be empirically supported, I here restrict my comment to *assumption 1*. To continue with the analogy between the normative theory of Kahneman, Wakker, and Sarin (1997) and prospect theory, note how *assumption 1* is very similar to the first psychological phase in prospect theory labelled as *editing/framing* (Tversky and Kahneman 1992, p. 299). In the editing/framing phase, the decision maker constructs a representation of the acts, contingencies, and outcomes that are relevant to the decision.

Tversky and Kahneman (1981) specify that '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' (p. 453 — my emphasis). One issue is that *assumption 1* also necessarily depends on personal and social characteristics of individuals, which can lead to very different perceptions of the good-bad scale among individuals.

Also, *assumption 1* obviously rules out any mental evaluation that goes beyond a 'good-bad' dimension. Unsurprisingly, the experienced utility criterion is then restricted to the evaluation of pain and pleasure in 'simple cases', e.g. a toothache, the taste of an ice cream flavour, the itch of a moskito bite, etc.[13]

2.3.4 AXIOM 4 (Interpersonal Comparability)

The comparisons of individuals experiencing different outcomes (e.g. a colonoscopy and the sensation of drinking tea) are ordinal, but the comparisons of individuals experiencing the same outcome (e.g. a colonoscopy or the sensation of drinking tea) are cardinal.

This axiom refers to the classical interpersonal comparisons of utility assumption that is subject to a long controversy in welfare economics. Because of the huge background on this historical controversy I obviously restrict my discussion to the arguments put forward by Kahneman et al.[14]

Recall that Kahneman and Varey (1991) argue that *adaptation* is one important reason which permits interpersonal comparisons of utilities. According to the authors, when two individuals are fully adapted to different levels of stimulation, they can be said to be matched in their absence of response to their states.

The other reason they bring about is if individuals' responses to stimuli differ in the same direction from their respective adaptation levels, those can be matched in signs, if not in magnitude. The main argument advanced by Kahneman, Wakker, and Sarin (1997) is that the functions that relate subjective intensity to physical variables are qualitatively similar for different individuals. This refers to what has been said previously about the non-impossibility of individuals to perceive the loudness of a sound similarly when it is of different pitch and timbre.

[13]The empirical support for the possibility of fitting various kinds of human sensation on a good/bad dimension is vast in psychology and is consequently outside the scope of the present chapter. I refer the reader to Kahneman (1999, pp. 7-9), who reviews the literature on this area of research, and who is enthusiastic about using the good-bad dimension for many situations — e.g. for the experiences of a straining runner and of a spectator watching a tragedy.

[14]See Fleurbaey and Hammond (2004) and Baujard (2017) for syntheses of this historical controversy.

Because the cardinal measurement of deviations in sign or in magnitude may not perfectly reflect adequate perception of feelings between individuals, it may explain why Kahneman, Wakker, and Sarin (1997) restrict cardinality between individuals in *one same situation* (e.g. a colonoscopy) but not in two different situations (e.g. a colonoscopy and carrying a heavy suit case).

I however suspect many scholars to find the assumption of interpersonal comparisons of utilities unsatisfying, even for individuals experiencing the exact same situation (e.g. a colonoscopy). The simple reason is that although the empirical arguments provided by Kahneman (2000) that the sign and intensity of current hedonic and affective experience is not essentially different from the standard psychophysical tasks of reporting colour or smell, many scholars would still believe that individuals may have incommensurable perceptions of pains, and some can react about the exact same pain in a very different manner than another. As Kahneman, Wakker, and Sarin (1997) put it,

> 'Of course, not all human pleasures and pains are biologically programmed in detail. Prior consumption experiences and various cultural and social influences can alter the hedonic value of stimuli, as when people learn to like coffee or chili peppers, develop a dislike for rich desserts, or acquire a passion for opera.' (p. 379)

The difference between tenants and adversaries of interpersonal comparisons of utilities would then be a matter of degree in terms of how far can we accept — biologically and sociologically speaking — individuals to perceive things similarly. This is a complex debate, which would require a large amount of studies before we can reach consistent knowledge of how individuals' perception differ in terms of pain and in terms of pleasure. As it may appear quite convincing that physical pain is perceived globally similarly among most individuals, it seems obvious that anything related to pleasure (which is at the end what the experienced utility criterion is designed for) is perceived differently among individuals.[15]

2.3.5 AXIOM 5 (Separability)

The order in which moment utilities are experienced does not affect total utility. That is, the contribution of an element to the total utility of the episode (or TEO) is independent of the elements that are preceded and followed it.

This axiom is perhaps the most important of the experienced utility criterion. Without it, the concept of total utility can simply not result from the summation of moment utilities, as total utility does not preserve the order in which moment utilities are experienced. Indeed, this axiom is needed to sum 'at will' all moment utilities of an episode of a TEO.

[15] A friend anaesthetist of mine told me a story he had at the emergency department about two patients, one Italian and one Vietnamese, which appeared to have two similar head injury diagnosis after they where taken care of by the medical staff. The Italian patient had a mild case and was screaming, while the Vietnamese patient had a severe case and was calm and silent. The medical staff was obviously not impressed by the behaviour of the Italian patient, that they are used to handle. Instead of judging the severity of patients' case by a subjective report based on a 0 to 10 pain-scale, they are accustomed to rely on symptoms that patients are asked to declare by answering several questions such as 'do you have nausea?', 'do you feel a contraction at the level of your temples?', etc. Eventually, the Italian patient appeared to be in so much pain that he was taken care of first by the medical staff. Whether he was actually more in pain than the Vietnamese patient (and if it is so, in which magnitude) or whether he was simply overreacting, remain mysteries.

Philosophically speaking, it is perhaps also the most contestable.

The axiom basically says that the sum of the experiences of playing a football game and having a beer is not affected by the order in which these two events are experienced. While it may appear obvious that having a beer after a football game is more enjoyable than the other way round, Kahneman, Wakker and Sarin (1997, p. 391) and Kahneman (2000, p. 192) reply to this kind of objection by emphasising that the episodes of a TEO that are to be evaluated are not *outcomes (or events)*, but *moment utilities associated with outcomes (or events)*. What does the distinction between outcomes and moment utilities associated with outcomes change the deal?

Recall that under Axiom 1 (inclusiveness), *all* the effects of the order of outcomes (or events) are already incorporated into moment utility. This means that when all moment utilities are summed, the social planner should not worry about the order in which those moment utilities are experienced because the information related to past and future events is already contained in the individual's moment utilities.

The issue is that by incorporating all previous and anticipated information in moment utility, one has specifically good reason to think that a total hedonic experience *will* be affected by the order in which these two moment utilities associated with events are experienced. In other words, it seems that physical events can be rearranged at will in time but once they are associated with a psychological affect, subjective experiences associated with events necessarily change.

As an illustration, consider the following two scenarios. Anticipating the enjoyment of having a beer after his football game (incorporation of information about anticipated utility), Jules attributes 6 hedonic state to the football game and 7 hedonic state to the beer he is now enjoying as a reward after decent effort (incorporation of information about past utility). *Scenario 1* therefore yields to a total utility of 13 hedonic states. Consider now a second scenario. Anticipating the episode of playing a football game while enjoying his beer, assume Jules attributes 5 hedonic state to the beer and -3 hedonic state to the unpleasant feeling of running on the pitch with a non-empty stomach. *Scenario 2* therefore yields to a total utility of 2 hedonic states.

Table 2.1: Hypothetical evaluation of hedonic scenarios

	$u(football)$	$u(beer)$	total utility
scenario 1: football then beer	6	7	13
scenario 2: beer then football	-3	5	2

If $13 \neq 2$, how can the order of these two episodes not affect the value of Jules' total utility? The counter-intuitive aspect of the *separability* axiom requires to discuss some of its underlying implicit assumptions. To 'appreciate the intuition' of this axiom, Kahneman (2000, p. 192) proposes the following thought experiment.

Assume an individual wins two unexpected prizes in a row: 500$ and 10 000$, then suddenly dies (or loses his memory). In evaluating the total utilities of both scenarios (*scenario 1'*: receiving 500$ then 10 000$; *scenario 2'*: receiving 10 000$ then 500$), *scenario 1'* would surely be preferable to him because the enjoyment of a smaller prize is greater when it comes first (equivalently, the enjoyment of the bigger prize is greater when it comes second).

Now let us imagine that all we know is that just before his sudden death (or amnesia), an individual had two pleasurable experiences, respectively $u(a)$ and $u(b)$, where $u(a) \gg u(b)$. Kahneman asks, 'would we still think that their order matters?', to which he replies that 'when outcomes are moment-utilities, there is no compelling reason to reject separability' (p. 192). This argument is however a bit fuzzy in the sense that it does not clearly specify what is at stake. Several points are worth being discussed.

First, this thought experiment makes it quite disturbing to perceive the relevance of the social planner's role in evaluating the individual's total utility. Those moment utilities experienced by the individual must matter to *the individual*, not to an external observer. But if the difference in total utility ultimately matters to the individual (and not the social planner), the difference between the value of the individual welfare function (or total utility) of *scenario 1'* and the one of *scenario 2'* should have reflected enough information to observe a salient magnitude between both individual welfare functions, just before the individual died.

As Kahneman (2000) seems to acknowledge it, as long as *scenario 1'* provides more total utility than *scenario 2'*, the first should be preferred to the second. This is true even if the difference in magnitude between the two total utilities is extremely small. Shall the order of moment utilities slightly disrupt the value of total utility, recall that the ethical premise of experienced utility states that the aim of the social planner is to maximise one's total utility (Section 2.1). Under such maximisation principle, it would then be sufficient to hold that the order does matter.

Second, and in relation with the first point, it is not that clear what the introduction of death (or amnesia) brings more to the argument if the evaluation of total utility of the individual is relevant *before* he dies (or get amnesic). Imagine you go to the restaurant. There is one scenario in which the order of the course goes normally, starting with the starter and ending with the desert. There is another scenario where the waiter brings you the desert at the beginning and the starter at the end.

What does you getting hit by a car when you get out of the restaurant brings up more to the evaluation of your concatenation of episodes at the restaurant from the social planner's viewpoint? The way I understand it, separability is relevant when the evaluation of one's total utility is made *after* the individual gets amnesic. For example, assume the lottery winner receives 500$, gets amnesic, then receives 10 000$. Would his total utility changed, had he received 10 000$, got amnesic, then received 500$? Presumably not.

For the sake of better practical appeal (it is rather uncommon that people get amnesic from one moment to another), let us take back the football-beer example. Assume 'football game' and 'beer' are not experienced at the same day but at two separate days (or even

at two separate weeks). In this case, it seems reasonable to hold that the order in which moment utilities are experienced does not affect total utility, simply because the distance in time between these two experiences is 'big enough' so that these experiences can be considered to be independent one from another.

Hence, the separability axiom seems to be reasonable under the condition that the distance between two temporally finite disjoint episodes/events is sufficiently big so that the subjective evaluation of one moment utility associated to an event does not affect the subjective evaluation of the other moment utility associated to another event. In other terms, the higher the distance in time between two episodes is, the more plausible it is to have two equal total utilities for both scenarios. There is however no imposed condition on the distance between two finite disjoint episodes in the definition of a TEO to construct total utility (see Appendix 2.A). If the present argument is judged to be relevant, *time-distance* may then be a required axiom to be added.

2.3.6 AXIOM 6 (Time Neutrality)

All moments are weighted alike in total utility. That is, the temporal distance between an outcome and its retrospective assessment is entirely irrelevant to its evaluation.

From a philosophical point of view (which this axiom clearly takes), time neutrality is the thesis according to which individuals should attribute no normative significance to the temporal location of their pleasure and pain (all else being equal). It is important to remind ourselves that total utility is always assessed *after* the moment at which the outcome is experienced. The idea is if the social planner takes a 'neutral' stance in summing all utility profiles, there is no apparent reason that he attributes more weight to one time at which one experience is evaluated by the individual instead of another.

To understand why Kahneman, Wakker, and Sarin (1997) and Kahneman (2000) make this normative judgement, consider first how individuals tend to weight time in decision utility and remembered utility. In decision-making, temporality *does* matter: economists assign to each intertemporal choice a discount factor, which captures the individual's patience. The more the outcome occurs late in time, the heavily the outcome is discounted. Remembered utility works the other way round: individuals' retrospective judgement tend to give more weight to the time at which the peak of pain is experienced and the final time at which the last intensity of pain is experienced (according to peak-end rule).

Kahneman (2000) however judges both decision utility and remembered utility to have a 'dubious normative status' (p. 193). According to the former, he brings up the classic argument in the literature of self-control failures that myopic preferences are normatively irrelevant (Thaler and Shefrin 1981; Laibson 1997) because they do not maximise total utility. According to the latter, the author judges that 'an experience that ended very badly could still have positive utility overall, if it was sufficiently good for a sufficiently long time' (p. 193).

A quick objection we can first make to this axiom is that attributing a 'neutral' value to time is far from being self-evident. Indeed, individuals may simply like to attribute

different weightings of time during the day because they have reasons to do so. For example, an individual who wakes up every morning to go to work may rationally think that his hedonic state of -2 does not have the same weight of his hedonic state of 7 when he goes back home. This is because the time associated with the negative feeling of making something unpleasant may not be perceived equivalently with the time associated with the positive feeling of playing with his cat after he gets back from work. The individual values the second activity much more than the former, and accordingly, cares less about the time of the day at which he makes something unpleasant.

He may also have the opposite reasoning, which is also consistent with time weighting. Consider that the pain he experiences by waking up every morning affects him more than the enjoyment of playing with his cat when he goes back home. This individual may have a negative remembered utility about his past TEO. Even if his total utility is positive, he may provide good reason not to want to repeat this TEO because he weights pain-time more than pleasure-time, to the point that he has a negative retrospective value of that TEO.[16]

This example may be receivable without further argumentation because it compares a pleasurable experience with a painful experience. No doubt individuals may value time differently in a TEO where both pain *and* pleasure are experienced, but how about in a TEO where either pain *or* pleasure is experienced? What I have to briefly discuss now is, *is it irrational not to consider time as being neutral?* Kahneman's normative stance about the relationship between time and rationality is in fact very similar to the one of Parfit (1984).[17]

To understand what is at stake, note that the example above says that the individual values more to play with his cat when he gets back from work rather than going to work because he *desires* one action more than the other. And it is because he desires one action more than the other that he has *reason* to weight time differently. Parfit (1984) disputes the Humean view, according to which rationality is only grounded on reasons to believe, and since a desire cannot be false (according to Hume), it cannot be open to rational criticism.

Parfit (1984) disagrees with this, arguing that rationality is not only grounded on reasons for *believing*, but also on reasons for *acting* (p. 120). According to Parfit (1984, p. 124), for temporal biases to be considered as normatively relevant (e.g. hyperbolic discounting), one must provide *reasons* for such behaviour.

> 'Someone is not irrational simply because he finds one experience more painful than another. But he may be irrational if, when he has to undergo one of these two experiences, he prefers the one that will be more painful. This person may be able to defend this preference. He may believe that he ought to suffer the worse pain as some form of penance. Or he may want to make himself tougher, better able to endure later pains. Or he may believe that by deliberately choosing now to undergo the worse of two pains, and sticking to this choice, he will be strengthening the power of his will. Or he may believe that greater suffering will bring wisdom. In these and other ways, someone's desire to suffer the worse of two pains may not be irrational.' (p. 123)

With this first point in mind, we can provide some reasons to question time neutral-

[16] This thought experiment implies that remembered utility has normative significance, which is the matter of discussion in Section 2.4.

[17] By rationality, I mean here 'what someone has reason to do'.

ity in the construction of total utility, and Parfit would perfectly agree with it. In the colonoscopy experiment an individual may prefer, for the reasons Parfit mentions (e.g. strengthening the power of one's will), to repeat the procedure which is more painful than the other, even if he actually remembers this experiment to be more painful. Now the main point is what if the individual does not have reason to do so, but simply has a desire for it?

Parfit (1984) answers this argument with another thought experiment of an individual who has '*future-tuesday-indifference*' (p. 124 — his emphasis). Imagine an individual who cares in a perfectly equal manner about the pain (or pleasure) that occurs to him in the future, except on Tuesday, where he does not care at all about the pain (or pleasure) he endures by then. To stick with only one hedonic state (pain), this means that 'he would choose a painful operation on the following Tuesday rather than a much less painful operation on the following Wednesday' (p. 124).

Parfit (1984) argues that preferring more pain to less simply because the agony will be on Tuesday '*is no reason*' (p. 124 — his emphasis). He then extends his argument, asking what would be the difference in principle with an individual who cares equally for everything that will happen to him within a year, but once a full year has passed, discounts by half the rest of his future. That is to say, this individual would rather choose e.g. two days of pain twelve months and one day from now rather than one day of pain twelve months from now. Parfit judges this kind of psychological rule to be simply arbitrary — along with the ones which discriminate between equal pleasures or pains:

> 'It is irrational to care less about future pains because they will be felt either on Tuesday, or more than a year in the future.' (pp. 125-126)

With Parfit's (1984) defence of time neutrality, we can first complete Kahneman's (2000) implicit argument that shall the individual have no reason about having this kind of preference, there is no point in considering each of her moment utilities extended in time as being non-neutral.

Second, if one agrees with Parfit (1984), one may need to justify this reason on something more than a belief. For example, to say that 'I prefer to give more value to the evening rather than the morning because I desire more what I do in the evening rather than what I do in the morning, even if I enjoy both equally' would be irrational according to Parfit if there is no reason associated with such desire ('I simply desire so but I cannot tell you why').

This second point is naturally a bit more complex because it gets quite philosophical. For practical purpose and for the respect of individuals' free will, do we need to provide reasons *why* individuals are irrational or not? In order to preserve their autonomy, we should obviously weight time at their will if such data is available. As economists are mostly concerned with cases where such data is unavailable, a philosophical assessment of this kind of ethical dilemma is perhaps needed.[18]

[18] For an extensive philosophical discussion of time neutrality, see Parfit (1984, pp. 170-177) and Brink (2011).

For lack of further philosophical assessment of time neutrality and of empirical knowledge about what individuals' preferences are, what we can nonetheless say is that discriminating between the values of different times in one period is no more demanding in terms of ethical judgements compared to the fact of not discriminating. I suspect Kahneman, Wakker, and Sarin (1997) and Kahneman (2000) to assume time neutrality for practical appeal: such assumption avoids them to invoke arbitrary criteria in order to discriminate between different times in one period. But discussing why (and how) time should be weighted inevitably leads us to complex philosophical assessments, as briefly discussed.

Consider now the last four axioms of experienced utility measurement. The external observer (or social planner) in the normative theory of Kahneman, Wakker, and Sarin (1997) has a knowledge about the use of the scale (because he is omnipotent). His task is to make comparative judgements about utility profiles. Those judgements must satisfy the following axioms in order to determine an equivalent relation between the original utility scale and duration.

2.3.7 AXIOM 7 (Concatenation of Neutral Utility Profiles)

The global utility of a utility profile is not affected by concatenation with a neutral utility profile.

This axiom considers neutral utility profiles, defined as profiles in which instant utilities are hedonically neutral (i.e. 'neither good nor bad.'). Discussing this axiom would lead us back to Axiom 3 (distinctive neutral point), so I move on.

2.3.8 AXIOM 8 (Monotonicity in Instant Utility)

Increases of instant utility do not decrease the global utility of a utility profile.

2.3.9 AXIOM 9 (Monotonicity in Total Utility)

Replacing one profile by another with a higher global utility increases the global utility of the concatenation of two utility profiles.[19]

Axioms 8 and 9 impose the requirement that a measure of instant utility should comprise all the information required for the determination of total utilities. That is to say, all the information that is needed to evaluate the goodness or badness of an episode must be incorporated in its utility profile. This means that any effect of previous or anticipated consumption on the utility of present consumption must be incorporated in the measure of instant utility. This basically refers to what has been said in Axiom 1 (inclusiveness).

[19] Axioms 7, 8 and 9 hold under the theorem which states that there exists a non-decreasing ('value') transformation function of moment-utilities, assigning value 0 to 0, such that global utility orders utility profiles according to the integral of the value of moment utility over time (Kahneman, Wakker, and Sarin 1997, p. 391). For the formalisation and proof of this theorem, see Kahneman, Wakker and Sarin (1997, pp. 400-402). The present section is bound to discuss the axioms of experienced utility, as they provide the necessary and relevant information about their theoretical issues. For further details about the technical construction of the experienced utility criterion, I refer the reader to Kahneman, Wakker and Sarin (1997, pp. 390-403) or to Appendix 2.A (without the theorems).

2.3.10 AXIOM 10 (Cardinality of Instant Utility)

The ordering of total utility of two utility profiles does not change if for both the instant utility level is increased by the same constant over an equally long period.

This last axiom is necessary for making cardinal measurement. As Kahneman, Wakker, and Sarin (1997) put it, 'the analysis becomes simpler if cardinal measurement of instant utility can be assumed, so that differences of instant utility are meaningful' (p. 392). Once cardinality is assumed, the social planner can rescale moment utility by its relation to duration. For example, if the social planner judges that one minute of pain at the hedonic state of -5 is equivalent with two minutes of pain at the hedonic state of -3, the social planner can rescale this original hedonic report by considering that -5 of the transformed scale is equivalent to the double of the original hedonic state of -3.

I have already discussed the properties of the original scale (Axiom 2 and 3) and the possibility of making interpersonal comparisons of utilities with cardinal measurement (Axiom 4). We can then move on to the last major theoretical issue of the experienced utility criterion (that it exclusively takes moment utility as its ethical content).

2.4 Moment Utility *versus* Remembered Utility

As mentioned in Introduction, the researcher who has perhaps contributed the most to the experienced utility research program — Daniel Kahneman — explicitly said to have abandoned such program because he might have not understand what happiness is about. Kahneman et al. initially considered that subjects in the experiments of Kahneman et al. (1993), Fredrickson and Kahneman (1993), Redelmeier and Kahneman (1996) and Schreiber and Kahneman (2000) made mistakes because they failed to accurately remember the moment utilities experienced during the episodes, which made them preferred the worst experience according to the logic of utility integration. Accordingly, Kahneman et al. took utility integration as a normative standard and considered failures of maximising moment utilities as *mistakes* (i.e. a prejudice against one's well-being).

The issue is that, as Kahneman-2018 acknowledges it, the logical rule of utility integration may not decently represent individuals' long term happiness. In fact, if we think that what matters is not happiness as 'living in the moment' but happiness as a durable mental state, then we may have better interest in defining happiness in terms of *remembered utility* rather than in terms of *experienced utility*.

Kahneman-2018 is sympathetic with the idea that what matters is not the utility experienced at the moment (as in Benthamian utilitarianism) but the *memory* individuals have about those experienced utilities — disregarding whether they reflect the highest intensity of pleasure (or the lowest intensity of displeasure) experienced during those episodes. The idea is that contrary to an experience which is enjoyed at the present moment, memory is a durable mental state, which stays in one's mind for a long time. In this sense, individuals choose their next vacation not as a *present experience* but as a *future memory*.

This could explain why individuals typically like to buy souvenirs or to take pictures

of their vacation. In doing so, they can enjoy their vacation not only at the moment they experience it but also for the rest of their life. This point echoes with one of the objections Kahneman and Sugden (2005) early stated towards the experienced utility criterion:

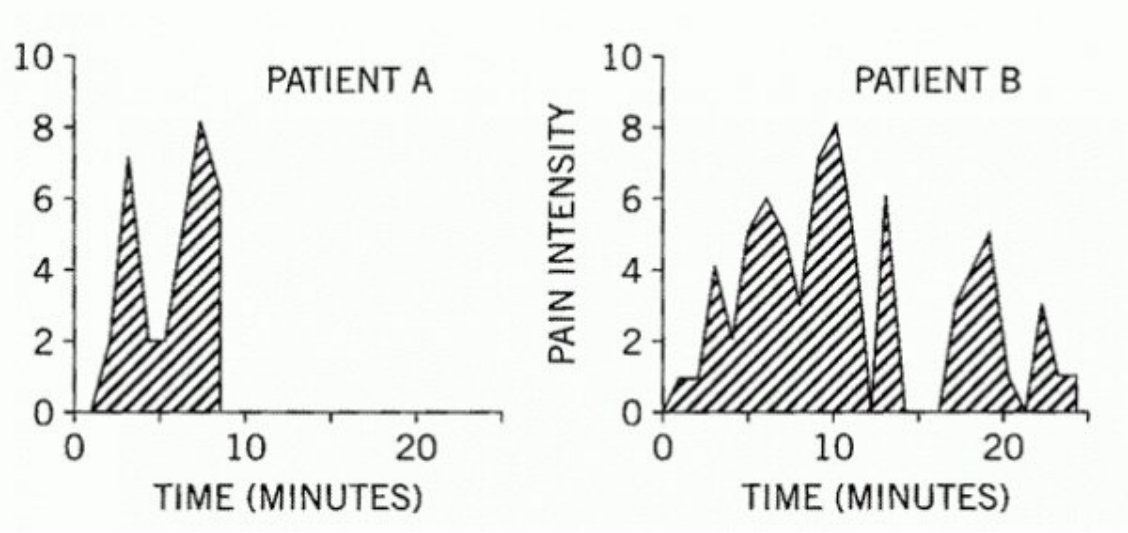
> 'It is possible to view life, not as a flow of pleasurable and painful experiences, but as the accumulation of a stock of good and bad memories. Because the mental representation of memory is more like a photograph album than a home movie — it is made up of discrete snapshots of "representative" moments — the life plan that maximises the integral of a person's happiness over time may not be the one that maximises the value of her accumulated stock of memories.' (p. 177)

The point is, if the logical rule of utility integration is considered to be unwarranted (given the way individuals represent the experience of their life taken as a whole), then the ethical premise of experienced utility falls apart. Considering that moment utility may not have more normative value than remembered utility, what are the implications for happiness measurement?

2.4.1 Remembered Utility Matters

First, we may need to reformulate experienced utility measurement with axioms that would not give normative value to *moment* utility but to *remembered* utility. That is, even if one's total utility is more painful than another (as in the cold-water experiment of Kahneman et al. (1993) or in the colonoscopy experiment of Redelmeier and Kahneman (1996)), the second one should prevail if most subjects hold the memory that it is less painful than the first one, *even if they actually experienced more total pain during the second experiment*. As an example, consider the following case.

Figure 2.2: Real-time recordings from two patients undergoing a colonoscopy.
Source: Redelmeier and Kahneman (1996).



The figure above displays the intensity of pain (y-axis) recorded each minute (x-axis) by two patients undergoing colonoscopy. Using my notations, the intensity of pain is measured by $\Psi = \{0, ..., 10\}$ and the outcome X (colonoscopy) associated with time $N = \{0, ..., 25\}$ is measured by the vector $x = \{x_1, ..., x_{25}\}$. Patient A experienced the short procedure (8 min) and Patient B experienced the long procedure (24 min).

According to peak-end rule, the total utility of the experiment with the added extra pain is remembered as less unpleasant than when no extra pain was added (specifically because this extra pain was *less unpleasant*). It is however clear that the total utility of the experiment with the extra pain is more unpleasant in terms of total utility than the short experiment.

Imagine, as in the experiment of hands submerged in cold water of Kahneman et al. (1993) that we, as policymakers, have to choose for a patient C who has undergone both types of colonoscopies one of the two colonoscopy to repeat. We have two alternatives: repeat the short experiment or repeat the long experiment. If Axiom 5 (time neutrality) and Axiom 6 (separability) hold, we can addition the three colonoscopy episodes in order to have two TEOs (temporally extended outcome) of the possible alternatives to evaluate: (i) the concatenation of 'short + long + short colonoscopies' episodes and (ii) the concatenation of 'short + long + long colonoscopies' episodes. Assigning a utility level of instant utility to each time point, we have the total utility profile of a TEO,

$$\sum_{1}^{3} \int_{0}^{N} u(x_i)dx$$

In terms of *experienced utility* (or utility profiles), we have,

$$\int_0^n u(x_i)dx + \int_0^{n+m} u(x_i)dx + \int_0^n u(x_i)dx > \int_0^n u(x_i)dx + \int_0^{n+m} u(x_i)dx + \int_0^{n+m} u(x_i)dx$$

where $m \in N$ represents the extra added pain (which equals to 16 min in the example above). According to the logic of utility integration, the concatenation of 'short + long + short colonoscopies' episodes dominates the concatenation of 'short + long + long colonoscopies' episodes. However, in terms of *remembered utilities*, it is the concatenation of 'short + long + long colonoscopies' episodes which dominates the concatenation of 'short + long + short colonoscopies' episodes (under the assumption that the extra added pain is *less unpleasant* than the short procedure, as in the example above). That is,

$$\int_0^n u^r(x_i)dx + \int_0^{n+m} u^r(x_i)dx + \int_0^n u^r(x_i)dx < \int_0^n u^r(x_i)dx + \int_0^{n+m} u^r(x_i)dx + \int_0^{n+m} u^r(x_i)dx$$

where $u^r(x_i)$ is the remembered utility of the patient at time i. The possibility of considering remembered utility as being more valuable than moment utility was actually already suggested by Redelmeier and Kahneman (1996), who concluded with the following words:

> 'For procedures where some pain is unavoidable, clinicians may need to decide whether it is more important to optimize patients' experiences or memories.' (p. 7)

In our example, patient C has a *false belief* that the total utility of 'short + long + long colonoscopies' is less unpleasant than the total utility of 'short + long + short colonoscopies'.

With respect to the discussion about Parfit's (1984) reasons for acting (Section 2.3.6), we have two possible schools of thoughts: the one which would say that any acting based on a false belief is necessarily an irrational behaviour (Kahneman et al.'s viewpoint), or

the other which would say that the individual's explicit verbal statement is not irrational in the sense that his belief about having less pain in the long experiment is *true to him*. Since the aim of the experienced utility criterion is not to maximise the social planner's well-being but the individual's well-being, we may judge that it is preferable to give normative value to remembered utility.

It appears that the example above also suggests a paradox regarding the theoretical construction of the experienced utility criterion. Recall that according to Nozick's (1974) argument (Section 2.2), it is not absurd to assume that tenants of the experienced utility criterion would say that what matters is not the individual's *true* beliefs about what he experiences, but what he *thinks he experiences*.

Shall the individual's brain be constantly manipulated by a benevolent scientist whose purpose is to maximise the individual's level of dopamine and to stimulate more regions in the brain where dopamine is active, it does not matter whether the individual's beliefs about experiencing this virtual world is false. In the same way, why should the individual's verbal statement in the real world not be taken as a false belief, which is, in the same logic, normatively relevant to the social planner?

The point of Kahneman et al. is that maximising remembered utility cannot be normatively relevant because it is considered to be 'biased': it gives more weight to the peak-time and the end-time of the procedure. However, the concept of remembered utility is, after all, only a matter of interpretation of the observer. Is remembered utility not a form of moment utility, which, as the definition of moment utility holds, incorporates the information of past and anticipated feelings in its evaluation?

Ultimately, the ethical premise of a normative approach that would give importance to individuals' remembered utility could be reformulated in terms of *negative utilitarianism* of *remembered utilities*: it is good to *minimise* the remembered disutility of one's suffering. That is,

Ethical premise (bis). *An individual's state of affairs is better than another if it has less remembered disutility than another*. Formally, let $x = (x_1, ..., x_n) \subseteq X$ be a realisable set of an individual's states of affairs (e.g. a consumption bundle, health states, sips of tea, etc.) and X be the set of outcomes. I denote by $i = \{0, ...n\}$ the index of time for each element of the vector x. For example, x_1 is one physical pain at time 1, x_2 another physical pain at time 2, and so on. $W(x)$ is an individual welfare function of the form,

$$W(x) = \int_0^n -u^r(x_i)dx$$

where $-u^r(x_i)$ is the individual's remembered disutility experienced at time $i = \{0, ..., n\}$ and $\int$ the integral of all utility profiles, which simply allows to have the total utility of this individual (here the total disutility of suffering). The remembered utility criterion is satisfied under the condition that,

$$W(x) > W(x') \implies x \succeq x'$$

2.4.2 Back to Decision Utility?

Second, if decision utility is mostly driven by remembered utility of a past episode — i.e. decision utility is an expression of an individual's preference for repeating one event over another — we are simply back to *decision utility* as the relevant normative criterion for normative analysis. In fact, that decision utility diverges from experienced utility is — strictly speaking — more a theoretical assumption rather than an observation supported by empirical evidence.[20]

Subjects in Kahneman et al.'s experiments are judged to make mistakes because of either retrospective judgement about their past experience, which showed that decision utility does not maximise experienced utility, or because of failure to predict their future (or anticipated) utility. That is, they make a mistake because of *fallible memory* and incorrect evaluation of *past experiences* or because of *wrong anticipation*.

But if this observation is not at hand (like in many circumstances where public policy applies), we need a counterfactual: what they would have done had they been able to maximise their experienced utility. In the many situations where a counterfactual preference is required to justify the experienced utility criterion (for lack of empirical evidence about the effect of a new policy implementation), how can we seriously assume that individuals' decision utility does not actually reflect their well-being? When observing an individual who has the choice between x and y and chooses x over y, we have simply no empirical evidence to claim that she would have been better off with y over x *at the time we are observing this*.

There is actually empirical evidence which disputes the common assumption that decision utility is fundamentally different than experienced utility (although we may accept the conceptual difference). Carter and McBride (2013) propose an empirical test of the similarity of shape and behaviour between the value function of prospect theory (which depicts individuals' choice), and the experienced utility function that is theoretically assumed in the normative theory of Kahneman, Wakker, and Sarin (1997).

Their empirical result can essentially be resumed in two lines: experienced utility is S-shaped (like the value function of prospect theory) when using the expectations and social comparison as the reference point, but is not always S-shaped when using past outcomes as the reference point. The result of their study lead them to suggest that decision utility and experienced utility are fundamentally related, although they are conceptually different.[21]

The empirical test of Akay, Bargain, and Jara (2017) in their paper named by the provoking 'Back to Bentham, Should We?' is even more concerning. Comparing British households' observed preferences with their reported subjective well-being, the authors found striking similarities on average between decision utility and experienced utility.

[20] See Kahneman, Wakker and Sarin (1997, p. 376), who justify the intuitive appeal of differentiating the two concepts of decision utility and experienced utility with the help of a thought experiment (and not empirical data).

[21] Note that Carter and McBride (2013) naturally acknowledge that the S-shape of both decision utility and experienced utility should be understood as one of the various possible shapes observed in a heterogeneous population (p. 14).

Their empirical study concludes that a majority of individuals made decisions that are actually consistent with the maximisation of their subjective well-being.

Eventually, Daniel Kahneman's journey in measuring experienced utility might end up to a useful wisdom for researchers interested in improving the methodology of subjective well-being measurement: the verbal statements at the end of each experience that violates monotonicity may in fact be the ones which can be considered to be normatively relevant. Does Kahneman-2018's acknowledgement about utility integration being a dubious normative standard yields to dispute the fundamental grounds on which the heuristics-and-biases program is based on: that individuals who deviate from the norms of rational choice make *mistakes*?

According to Kahneman, Wakker and Sarin (1997, pp. 377, 395) and Kahneman (1999, p. 20), empirical evidence showed that individuals already have the ability to maximise the utility they store in their memory (i.e. individuals maximise their remembered utility). When this empirical evidence is combined with the ethical stance that remembered utility may actually be what matters (like Kahneman-2018 states), we are simply back to *decision utility* (and thus observed choice) as the proper normative criterion for public policy.

2.5 Conclusion

In the present chapter, my aim is to provide an up-to-date assessment of the whole program of experienced utility measurement after the reconsideration of Kahneman-2018. My analysis follows four steps. I first provide a literature review of the program of Kahneman et al. I then consider several issues of Benthamian hedonism for public policy. Then, I provide a philosophical discussion of all the axioms of experienced utility measurement. Eventually, I aim to persuade my reader that measuring experienced utility is based on a misconception of happiness that economists and policymakers have good reason to stay away from. The highlight of the chapter is that all the methodological and theoretical issues discussed throughout my analysis provide economists and policymakers strong support for endorsing alternative measures of happiness that do not aim at maximising pleasure, but which are grounded on perhaps better objective conceptions of what makes the good life.

As an illustration of what those alternative measures might be, consider Kahneman-2018's new distinction of two concepts of happiness: (i) the feeling of enjoyment an individual has at the moment, which is related to the experiences she has at the moment (*moment utility*), and (ii) the feeling related to social yardsticks such as achieving goals and meeting expectations, which is based on comparisons with other people (*life satisfaction*). In Kahneman-2018's new terms, objective happiness is more about life satisfaction in terms of social life, i.e. the relationship with the company of others (partner, friends, family and colleagues) rather than the maximisation of pleasurable moments.

It seems not absurd to consider that Kahneman-2018 has switched from Benthamian hedonism to *Aristotelian eudomonism*. In contrast with hedonism (in greek, *hedone* for pleasure), eudaimonism (in greek, *eudaimonia* for happiness) does not put the satisfaction of pleasure at its central ethical principle. It instead considers a broader perspective of

what makes the good life, typically friendship and the participation in civil or political life (Aristotle -350 [2009]). In other words, according to Bentham pleasure is identical with happiness (and the goal of life is to produce the greatest happiness for the greatest number), while according to Aristotle pleasure is not identical with happiness but can be either a component, a process or a by-product of it.[22]

The main issue of experienced utility measurement seems to be that utility integration invokes a conception of objective happiness that is paradoxically based on an extremely *subjective* informational basis of happiness (i.e. moment utility). Recall that under Axiom 1 (Section 2.3.1), only hedonic states are normatively relevant, and nothing else. But considering Kahneman-2018's statement that what matters is life satisfaction rather than moment utility, economists and policymakers may want to promote measures of happiness that do not depend on individuals' subjective perception. Instead, they may want to promote 'authentic' objective features about what makes the good life such as health or friendship (as in Aristotelian terms). By 'authentic' I mean that the content of such objective measure would not be a subjective feeling that is up to strong variations among individuals. On the contrary, it would be something stable on which individuals could perhaps more consensually agree about, e.g. the opportunity to live a life where basic human needs such as health, education and friendship are fulfilled.

Thus the capability approach (Sen 1985; Nussbaum and Sen 1993; Nussbaum 2000) is perhaps the best way to take this (already taken) route. Capability is defined as what people are capable of achieving based on the opportunities and living conditions afforded then. In this normative approach, what makes the good life is not merely defined in terms of a subjective perception like in the experienced utility criterion, but in terms of essential human needs. Ten 'central human functional capabilities' are offered by Nussbaum (2000): life; bodily health; bodily integrity; senses, imagination and thought; emotions; practical reason; affiliation; other species; play; control over one's environment (pp. 78-80). Notice that all these criteria of what makes the good life are actually *opportunities* to do something, e.g. to live a normal life, to access to appropriate level of housing, etc. These criteria are potentially far more likely to reach a consensus about what makes the good life among all living populations than subjective rankings in terms of pain and pleasure. The reason is that subjective rankings in terms of pain and pleasure are likely to be more sensitive to personal/social norms and personal/social comparisons. Consequently, human capabilities perhaps represent more appealing characteristics of what *objective* happiness actually is.

In his *Nicomachean Ethics*, Aristotle (-350 [2009]) defined happiness as the activity chosen for its own sake by a morally serious and virtuous person. According to the philosopher, happiness is a harmonious psychological state in which the individual lives a virtuous life that includes not only the seek of pleasure, but more importantly an excellent

[22] See Nussbaum (2007) for a philosophical comparison between the ethics of Aristotle and Bentham. Nussbaum particularly studies the case of J.S. Mill, who according to the author aims at combining Benthamian and Aristotelian conceptions of happiness. The thesis of the author is that 'despite Mill's unfortunate lack of clarity about how he is combining the two conceptions, he really does have a more or less coherent idea of how to combine them, giving richness of life and complexity of activity a place they do not have in Bentham, but giving pleasure and the absence of pain and depression a role that Aristotle never sufficiently maps out' (p. 172).

trait of character such as being 'fair', 'wise' and 'honest'.[23] But in order to realise this psychological state, one should have access to resources that actually gives her opportunity to achieve this state of mind. The bottom line is if experienced utility measurement is flawed, policymakers and economists may seriously consider eudaimonistic conceptions of happiness that perhaps better capture what objective happiness actually is. I then suggest a new slogan for researchers interested in dropping off the last vestiges of the experienced utility criterion for happiness measurement (in memory of Samuelson): 'Back to Aristotle? Exploration of Objective Happiness'.

[23]Virtue ethics (often presented as neo-Aristotelian ethics) is considered to be one among the three main theories of ethics alongside deontologism and consequentialism. The concept of 'virtue' is extremely rich. For obvious reason it cannot be explained here. For a comprehensive review, see Hursthouse (2016).

2.A Glossary of the Experienced Utility Criterion

Decision utility is the weight of a decision inferred from choice, which is in turn used to explain choice. For any given alternative, e.g. 'drinking your tea' or 'reading this chapter', you have an assigned numerical value (either positive or negative) that represents your decision utility. Formally, let $X = \{x, y\}$ be the set of alternatives, where $x = (x_1, ..., x_n)$ is the vector that corresponds to the activity of drinking your tea. For example, x_1 is 'one sip of tea', x_2 is 'another sip of tea', and so on. Let also $y = (y_1, ..., y_n)$ be the vector that corresponds to the activity of reading this chapter, e.g. y_1 is 'reading one piece of this chapter', y_2 is 'reading another piece of this chapter', and so on. Let $u : X \longmapsto \mathbb{R}$ be a utility function. If $u(x) = 4$ then the numerical value of 4 is your decision utility of choosing x. If $u(y) = 3$, then the numerical value of 3 is your decision utility of choosing y. Because this numerical value has no psychological meaning in terms of hedonic state, we will here only account for the set X, not $\mathbb{R}$.

Remark 1. Like in standard microeconomics, the utility function $u : X \longmapsto \mathbb{R}$ is a way of assigning a number to realisable alternatives such that more preferred alternatives get assigned a larger numerical value than less preferred alternatives. But the numerical value is here only relevant to allow for an ordinal ranking of decision utilities. It does not express the psychological intensity of the alternative chosen (contrary to experienced utility defined below).

Experienced utility is the hedonic state experienced in doing (or choosing) something. For any given alternative, e.g. 'drinking your tea' or 'reading this chapter', you have an assigned hedonic state (expressed in a numerical value), which describes your psychological intensity. Your experienced utility is high if it pleases you or low if it bothers you. Formally, let X be the set of alternatives and $u : X \longmapsto \Psi$ a utility function, where $\Psi = \{-10, ..., 10\}$ is the set of hedonic states (-10 for the less pleasant feeling, 10 for the most pleasant feeling). Assume $u(x) = 8$, then the numerical value of 8 represents the experienced utility of choosing x. If alternatives are of similar nature, e.g. 'one sip of tea' and 'another sip of tea', then cardinality applies (Axiom 4 [Section 2.3.4]). That is, let x_1 be 'one sip of tea' and x_2 be 'another sip of tea'. If $u(x_1) = 6$ and $u(x_2) = -3$, then the first sip of tea has exactly 9 more hedonic intensity than the other sip of tea.

Moment (or instant) utility is an attribute of experience formulated into a hedonic value, which is experienced at the present moment. It is the valence (good or bad) and the intensity (mild to extreme) of current affective or hedonic experience. For example, the enjoyment (or suffering) you are having in reading this chapter right now is of a given intensity, which only depends on your personal evaluation (e.g. you really like it, like it, are being indifferent, do not like it, do not like it at all, etc.). Moment utility is measured by asking subjects to evaluate their happiness on a hedonic scale (e.g. -10 the lowest hedonic state, 10 the highest). The set of moment utility (or hedonic states) is denoted by $\Psi = \{-10, ..., 10\}$.

Remark 2. As Kahneman, Wakker and Sarin (1997, p. 398) put it, the set of moment utility Ψ should include the neutral value 0. This is because negative feelings should be distinguished from positive feelings and to allow for cardinal measurement of moment utility on a ratio scale (Axiom 10 [Section 2.3.10]).

An episode is a connected time interval described by its temporal coordinates. For example, from the time you started reading this chapter until the time you are currently reading this, one episode has passed. Formally, let $[B, E[\in N$ be a time interval that contains all time points relevant to the analysis and let X be the set of outcomes. An episode is a function $f : [b, e[\longmapsto X$, for $B \leq b$ and $e \leq E$.

Remark 3. All time intervals are assumed left-closed and right-open because the union of episodes should not include twos slice times of different episodes (see below).

A temporally extended outcome (TEO) is a group of one or more temporally finite disjoint episodes. For example, from the time you started reading this chapter until the time you reached the previous definition and now this other definition, *two* episodes have passed. A TEO is simply the union of two (or more) separated episodes. Formally, a TEO is a mapping from a finite disjoint union of subintervals of the time interval $[B, E[$ to the set of outcomes X. That is, $f : [b, e[\cup [b', e'[\longmapsto X$ is one TEO, $f : [b, e[\cup [b', e'[\cup [b'', e''[\longmapsto X$ is another TEO, and so on. We can denote the general definition of a TEO by $f : 2^{[B,E[} \longmapsto X$, where $2^{[B,E[}$ is the set of all possible collections of subintervals in $[B, E[$.

A Utility profile of a TEO (or simply utility profile) is a function which assigns a level of moment (or instant) utility to each time point. Informally, we can interpret it as an extensive definition of moment (or instant) utility by introducing time as an explicit variable, thus allowing moment utility to fit in any temporality (either a time slice, an episode or a TEO). For example, the enjoyment (or suffering) you are having in reading this chapter (in a given intensity) can be represented at time 1, time 2, etc. Formally, a utility profile is a function $u : 2^{[B,E[} \longmapsto \Psi$, with $[B, E[$ the set of slices in time.

Remark 4.1. In order to keep the standard notation '$u(x)$' in the text, I however consider the summation of *experienced utility* (and not of utility profiles) to be the informational basis of total utility. This is far from absurd, since we only have to index experienced utility with time to have an equivalent notion with utility profile (although both mathematical objects are obviously different). That is, we can denote a utility profile by $u(x_i)$, where $i = \{0, ..., n\}$ is the index of time. I judge this simplification to be useful in order to avoid entering into technical details that are not fundamentally important to the global discussion of this chapter.

Remark 4.2. Kahneman, Wakker and Sarin (1997, p. 398) actually distinguish a *dated utility profile* from a *neutral utility profile*. The former defines the general concept of utility profile. The latter allows for a technical transformation so that some specific level of instant utility experienced at a given slice of time yields the same amount of instant utility at another slice of time, *independent of when it happens in history* (Axiom 5 [Section 2.3.5]).

Total utility is the addition of all utility profiles of an episode or TEO under the assumption that Axioms 1, 2, 5 and 6 of utility integration hold (Section 2.3). For example, from the time you started reading this chapter until the time you are currently reading this, you had two sips of your tea. The addition of the two utility profiles 'first sip of tea' and 'second sip of tea' is described by the total utility of the time interval in which you made these two things separately. Formally, let $u(x_1)$ be one utility profile at time

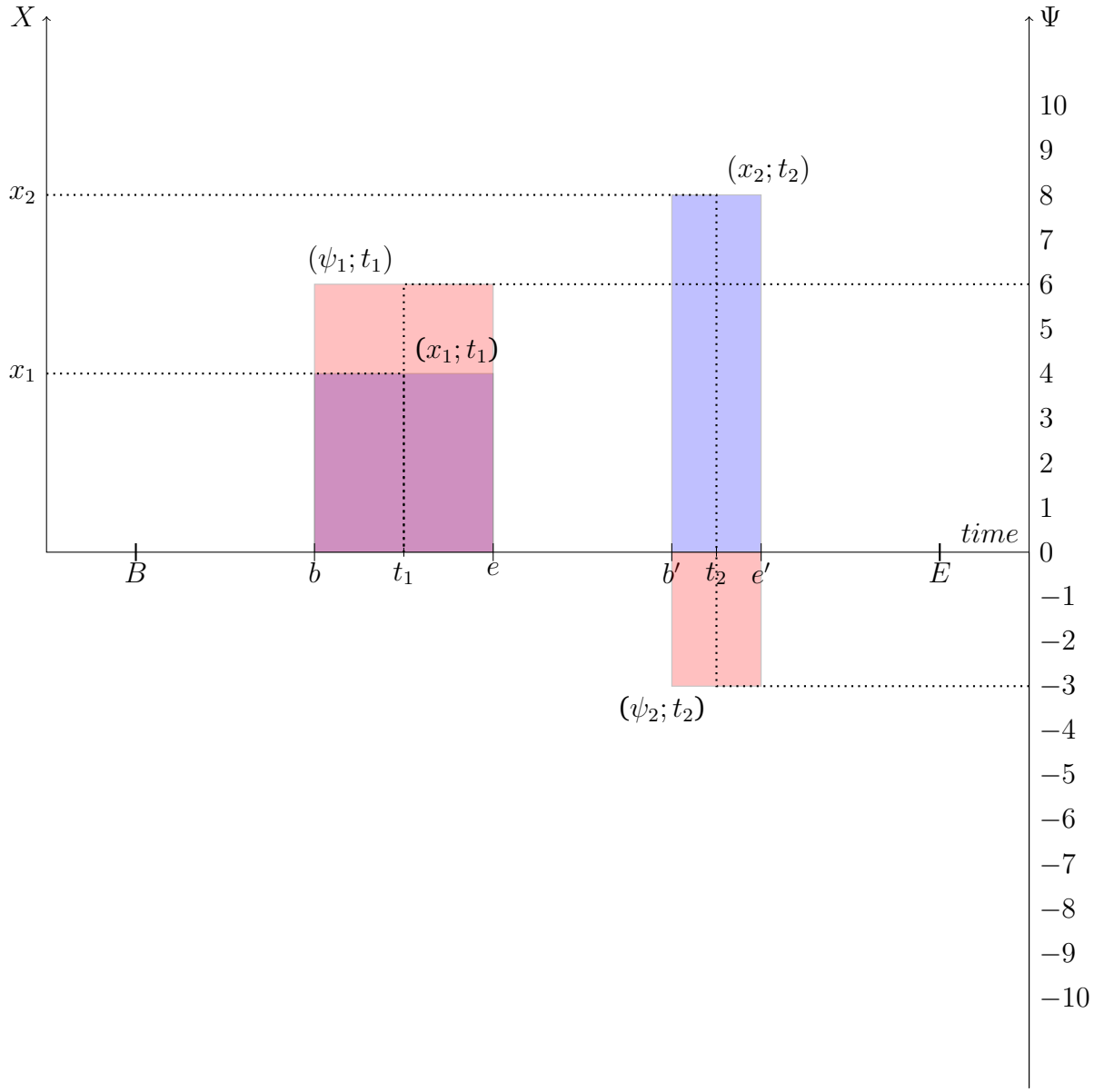
1 and $u(x_2)$ another utility profile at time 2. $W(x) = u(x_1) + u(x_2)$ represents the total utility of experiencing x_1 at time 1 *and* x_2 at time 2. From a social choice point of view, total utility is nothing more than an objective function a benevolent social planner aims at maximising. With the simplified notation I propose, total utility can be denoted as $W(x) = \int_0^n u(x_i)dx$.

Remembered utility is an individual's own global retrospective evaluation of a past experience, either represented in an episode or a TEO. For example, what you previously read of this chapter is a memory of a past experience. The evaluation you have about this past experience (either positive or negative) is your remembered utility of that experience. Formally, let X be the set of alternatives and $u^r : X \longmapsto \Psi$ a remembered utility function, where $\Psi = \{-10, ..., 10\}$ is the set of hedonic states (which represents the hedonic -10 to 10 scale). If we empirically observe (through your verbal statement) that $u^r(x) = 8$, then the numeral value 8 represents the remembered utility of thinking about the past experience of x. If we empirically observe (through your verbal statement) that $u^r(y) = 1$, then the numeral value 1 represents the remembered utility of thinking about the past experience of y. We can denote remembered utility by $u^r(x_i)$, where the superscript r stands for 'remembered' and where the subscript $i = \{0, ..., n\}$ stands for the time at which the individual thinks of her past experience.

Predicted (or anticipated) utility is a belief about future experienced utility. For example, the activity you are thinking of doing after you will be done reading this chapter is your predicted utility, also quantified in terms of hedonic states. Formally, the representation of predicted utility is exactly the same as remembered utility, except that since evaluation are not about *past* but *future* events, we can denote $u^p(x_i)$ for 'predicted'.

The following graph provides a visual representation of the relation between time (N), outcomes (X) and hedonic states (Ψ). 'Time' and 'hedonic states' are quantitative variables, while 'outcomes' is a qualitative variable (meaning it does not have a numerical value).

Figure 2.3: Graphical representation of experienced utility measurement



The x-axis represents the time variable N, to which each slice of time or interval belongs to. The time interval $[B, E[$ contains all time points relevant to the analysis, e.g. the evaluation of your evening. The intervals $[b, e[$ and $[b', e'[$ contained in $[B, E[$ are two distinct episodes, e.g. $[b, e[$ represents one hour and $[b', e'[$ represents thirty minutes. The finite disjoint union of $[b, e[$ and $[b', e'[$ which maps to a set of outcome X is a TEO. Visually, it is represented by the blue area, where $\{x_1, x_2\} \in X$.

The y-axis on the left represents outcomes (a qualitative variable), where x_1 and x_2 are two outcomes, e.g. 'first sip of tea' and 'second sip of tea'.

The y-axis on the right represents the hedonic scale $\Psi = \{-10, ..., 10\}$. The highest the value, the more enjoyable the experience is and conversely. The experience of one or several outcomes (e.g. drinking one or several sips of tea) is represented by a utility profile. A utility profile is a function $u : 2^{[B,E[} \longmapsto \Psi$. In the present illustration we have two utility profiles: $f : [b, e[\longmapsto \Psi$ and $f : [b', e'[\longmapsto \Psi$. Visually, a utility profile is

represented by the red area, where $\{\psi_1, \psi_2\} \in \Psi$. For the sake of illustration, the outcome x_1 gives you a hedonic feeling of 6 (because the tea is warm), while the outcome x_2 gives you a hedonic feeling of -3 (because the tea is now cold).

On the assumption that Axiom 5 (separability [Section 2.3.5]) and Axiom 6 (time neutrality [Section 2.3.6]) hold, we can represent the sum of two utility profiles into a total utility function of the form $f : [b, e[\cup[b', e'[\longmapsto \Psi$, or with the simplified notation I suggest, $W(x) = \int_0^n u(x_i)dx$. Since there are here only two experienced outcomes at two different slices of time, we have $W(x) = u(x_1) + u(x_2)$. The goal of the social planner is to maximise $W(x)$.

Remark 5. Again, representing total utility in terms of utility profiles would have required to write $W(n) = \int_B^E u(2^{[B,E[})dx$. This notation is avoided for two reasons. First, and as previously said, the notation $u(x_i)$ simplifies things. That is, I simply consider in the text that x is an element included in the two nested sets X and $[B, E[$. To make things even simpler, I use in the text the set of time N instead of $[B, E[$, where i is the index which captures each time slice. Second, Ψ actually depends on X, as in the definition of experienced utility. But the relationship between N, X and Ψ is quite peculiar. As Kahneman, Wakker and Sarin (1997, p. 398) put it, 'the instant utility at a time point depends on the outcome associated with that time point, but also on outcomes associated with other time points.' Under Axiom 1 (inclusiveness [Section 2.3.1]), not only a moment utility includes the present hedonic feeling ψ_i of doing x_i, but also of thinking about x_{i-1} being done and of anticipating doing x_{i+1}. In other words, all the information about experienced and anticipated outcomes are already included in ψ_i. This psychological phenomenon is however hard to represent graphically. It cannot be represented in a three-dimensional graph because the relation between variables N, X and Ψ is not a one-to-one mapping. That is to say, one element of X at time i maps to one element of Ψ at time i, but one element of Ψ at time i maps to several elements of X at different times, e.g. $i - 1$ and $i + 1$. Mathematically, it would also require to specify the particular relation between X and Ψ. Since ψ_i not only depends on x_i but also on x_{i-1}, x_{i+1}, and so on, we should technically denote $\Psi_i = f(X_i, X_{i'}), \forall i' \in 2^{[B,E[} \neq i$.

Remark 6. The graph provides a visual representation of the theoretical discrepancy between decision utility and experienced utility. But since the set of outcomes X is a qualitative variable, the distance between $(x_1; t_1)$ and $(\psi_1; t_1)$, and the distance between $(x_2; t_2)$ and $(\psi_2; t_2)$ are meaningless. I say 'theoretical' because the empirical studies of Kahneman et al. only show a discrepancy between predicted utility and experienced utility (Kahneman and Snell 1990, 1992) and between remembered utility and experienced utility (Kahneman et al. 1993; Fredrickson and Kahneman 1993; Redelmeier and Kahneman 1996; Schreiber and Kahneman 2000). But whether decision utility and experienced utility are fundamentally distinct is yet another question (Section 2.4.2).

CHAPTER 3

The Real Problem of the Reconciliation Problem: What Is a Good Normative Criterion?

Abstract

The issue of finding a normative approach consistent with behavioural economics has been coined as the 'reconciliation problem' by McQuillin and Sugden (2012). Yet as the authors put it, 'nothing remotely like a consensus has been reached about how the problem is best tackled' (p. 554). This chapter is an attempt at suggesting a consensus on how the reconciliation problem can be best tackled. It aims at giving it a unified structure so that some important normative criteria proposed as responses to the reconciliation problem — *experienced utility*, *true preference*, *choice-basis* and *opportunity* — can be assessed by a simple framework. After surveying the methodological and theoretical issues associated with each of these normative criteria, I propose that solving the reconciliation problem requires to consensually agree about value judgements that economists inevitably have to make regarding *what a good normative criterion is*. Accordingly, I suggest three essential requirements that a good normative criterion needs to satisfy. A criterion is *general* if it can be applied to a wide range of choice situations. It is *ethical* if it can capture the many different aspects of life that individuals can find valuable. Lastly, it is *practical* if it can measure individuals' states of affairs, and which measurement is relatively consensual. The result is that none of the *experienced utility*, *true preference*, *choice-based* and *opportunity* criteria satisfy all three requirements. This leads me to suggest avenues of future research on proposing alternative normative criteria that could potentially better satisfy these three requirements. These alternative normative criteria are the *virtue ethics* criterion and the *meaning* criterion.

Acknowledgement. Version August 2020. A preliminary draft of this chapter entitled 'Reconciling Normative and Behavioural Economics: A Critical Survey' was presented at the 2019 conference on Public Economic Theory in Strasbourg and at the 2019 Happiness Economics and Interpersonal Relations conference in Rome. Fragments of what constitute the present chapter were also presented at the 2018 Economic Philosophy conference in Lyon and at the 2018 European Society for the History of Economic Thought summer school in Argalasti. I thank the audience for helpful comments. I also thank Jean-Sébastien Gharbi and Cyril Hédoin for careful reading. All mistakes remain my own.

3.0 Introduction

The reconciliation problem is characterised as the problem of finding a way to do normative economics when individuals' preferences are incoherent. The problem is if the assumption of coherent preference breaks down, the standard normative criterion of preference-satisfaction is no more relevant to indicate what makes individuals better off. There is however a conceptual fuzziness about what sort of consensus the reconciliation problem can reach on how the problem is best tackled. This point is well emphasised by McQuillin and Sugden (2012, p. 554), who do not provide a clear-cut framework of the reconciliation problem except of course giving it its main guiding line. In addition, the reconciliation problem is interpreted by many perspectives in the literature, which obviously depend on the various researchers' interests.

To name a few perspectives, some authors focus on the identity problems of the individual (Lecouteux 2015a; Dold and Schubert 2018), others on various issues of libertarian paternalism (Grüne-Yanoff 2012; Guala and Mittone 2015; Hédoin 2015; Sugden 2017b; Hands 2020), others on quantitative techniques for eliciting individuals' true preferences (Bleichrodt, Pinto, and Wakker 2001; Pinto-Prades and Abellan-Perpiñan 2012), others on questioning the soundness of the true preference assumption (Sugden 2015; Lecouteux 2016; Infante, Lecouteux, and Sugden 2016a), others on theoretical models for identifying mistakes (Köszegi and Rabin 2007, 2008; Bernheim 2016), yet others on the relevance of the preference-satisfaction criterion itself (Hausman 2012, 2018; Hédoin 2017). In the same way Bernheim (2016) introduces his 'unified' approach to behavioural welfare economics, one may say that 'reading the literature, one can take the impression that the [reconciliation problem] has become a bit of free for all' (p. 13).[1]

The aim of this chapter is to give the reconciliation problem a unified structure so that researchers interested in this topic are invited to debate about what fundamentally matters to the reconstruction of normative economics when individuals have incoherent preferences. By 'unified', my aim is to propose a framework on which economists can consensually agree about what it takes for a normative criterion to be a 'good' normative criterion. The idea is once we consensually agree on some basic requirements that a normative criterion should satisfy, all we have to do is to assess whether one criterion satisfies those basic (or fundamental) requirements. The idea is if all requirements are satisfied, the criterion is judged to be 'good'. Otherwise it is not.

Fundamentally, normative economics is based on the use of normative criteria to evaluate individual or social states of affairs, and then to recommend/prescribe public policy based on such evaluation. We may then see the reconciliation problem to be essentially about *normative criteria*. The point is if there is a reconciliation problem to be solved, one first needs to provide an answer to the simple and fundamental question of '*what is a good normative criterion?*' In other words, the issue that there is yet no consensual approach about how the problem is best tackled can be formulated as follows. *What basic requirements should a normative criterion satisfy so that it is considered to*

[1] In the original quote, Bernheim (2016) refers to behavioural welfare economics, not to the reconciliation problem. But if behavioural welfare economics is one answer (among others) to the enquiry of finding a normative approach consistent with behavioural economics, it is only a subset of the reconciliation problem.

be a candidate for 'solving' the reconciliation problem? My suggestion is that these basic requirements can take the following three forms. A normative criterion is judged to be 'good' if,

1. *General requirement.* It can be applied to a wide range of choice situations.
2. *Ethical requirement.* It can capture the many different aspects of life that individuals can find valuable.
3. *Practical requirement.* It can measure individuals' states of affairs using a relatively consensual measure of what makes individuals better off (to be discussed below).

By definition, a normative criterion is a rule that tells us whether one outcome is better than another. We have then three particular requirements which account for three particular problems: 'better when?', 'better to what?' and 'better how?' Answering the first question implies to have an overall idea of the domain in which a given normative criterion applies. To define the 'normative-relevant' domain is necessary because we need to delimitate the boundaries of the normative relation R, i.e. what it can and cannot evaluate.[2] Answering the second question implies to have an ethical judgement over the normative-relation 'better than'. To define a normative relation R, outcomes x and y, and to say that $x\ R\ y$ means that 'x is better than y' is mathematically purposeful, but meaningless if we do not define the content of this normative relation. Answering the third question implies to have a measure of this ethical content that allows to practically evaluate individuals' states of affairs, without which no evaluation would be possible.

In sum, the general requirement is the the ability of a normative criterion to apply to a wide range of choice situations (the *scope* of the normative-relevant domain); the ethical requirement is the ability of a normative criterion to 'cut up the world', i.e. to judge what situation is considered to be better than another regarding the interests of individuals (the *content* of the normative-relevant domain); the practical requirement is the ability of a normative criterion to measure individuals' states of affairs (the *measurability* of the normative-relation). In the present survey, I argue that these requirements provide solid grounds for paving the way towards a unified account of the reconciliation problem. As I recognise this work to be only an attempt to unify the reconciliation problem — and as a result, that it will obviously not be 100% consensual — the important issues associated with the reconciliation problem can be treated step by step. That is, *attempts* to unify it would facilitate better perspectives of future research.

The rest of the chapter is organised in five sections. Section 3.1 presents the reconciliation problem. Section 3.2 provides a critical review of the main normative criteria offered in the literature: the *experienced utility*, *true preference*, *choice-based* and *opportunity* criteria. These normative criteria can be seen as direct responses to the reconciliation problem in the way they account for individuals' incoherent preferences in different manners. Section 3.3 develops the content of the *general*, *ethical* and *practical* requirements. Section 3.4 then assesses the *experienced utility*, *true preference*, *choice-based* and *opportunity* criteria with respect to the three requirements I propose. As we shall see, if

[2] I deliberately do not use the usual vocabulary of 'welfare-relevant domain' or 'welfare-relation' because a comparison is not necessarily based on *well-being* (or welfare, or individual utility). Otherwise, normative economics would be restricted to a narrow range of applications.

these requirements are accepted then none of these normative criteria satisfies them all. This leads me to suggest avenues of future research on proposing alternative normative criteria that could perhaps better satisfy these three requirements. I end up discussing those alternatives in Section 3.5.

3.1 The Reconciliation Problem

Behavioural economics has originally been developed on empirical and theoretical grounds in order to improve our understanding of decision-making. But from the 1990s it sparks an interest in normative analysis by questioning the relevance of the preference-satisfaction criterion of standard normative economics. Standard normative economics — which archetype can be labelled as standard welfare economics — conventionally uses preference-satisfaction as a reliable proxy of well-being and assumes that individuals have coherent preferences. The satisfaction of individuals' coherent preferences is then considered to be the proper normative criterion for evaluating individuals' states of affairs. However, a large body of empirical evidence document cases where individuals have incoherent preferences (Tversky and Thaler 1990; Kahneman and Tversky 2000; DellaVigna 2009). This is not only concerning for positive economics, which for around three quarters of a century has been grounded under the assumption of rational choice. It also impacts normative economics in the way that *observed* preferences are to be disentangled from *normative* preferences (which represent individuals' actual interests).

In brief, when individuals' preferences are incoherent due to numerous reasons (non-Bayesian updating, framing, non-exponential discounting, intransitivity, *status quo* bias, violation of dominance, etc.), economists do not have a stable normative concept so that they can evaluate individuals' states of affairs. To put it differently, if the assumption of coherent preference breaks down, the standard normative criterion of preference-satisfaction is no more relevant to indicate what makes individuals better off.[3] The problem of finding a normative approach consistent with the observation that individuals have incoherent preferences has been coined under the 'reconciliation problem' by McQuillin and Sugden (2012). As the authors put it, the concept of preference-satisfaction in standard welfare economics used to be relatively unproblematic as long as we assumed individuals to have coherent preferences. In fact, preference-satisfaction could be justified under the following three interpretations.

- *Happiness interpretation*. Individuals seek to maximise their happiness.
- *Well-being interpretation*. Individuals act according to their well-being (which does not necessarily reduce to their happiness).
- *Freedom interpretation*. Individuals act with respect to the freedom of making their own decisions (in line with the neoclassical consumer sovereignty principle).

[3] By 'no more relevant' I refer to situations where individuals are unlikely to have coherent preferences. There is of course the argument that if decision tasks are well-specified and are repeated a sufficient number of times — the 'discovered preferences' hypothesis Plott (1996) — the repetition of the game in a market environment enables individuals to get more experience so that they can *in fine* 'discover' their coherent preferences. What can nonetheless be said is that the documentation of incoherent preferences for decisions which are not sufficiently repeated — and actually also for some which are sufficiently repeated *and* made by market experts (Dhami 2016, p. 1449) — is significant enough to think about how normative economics should handle incoherent preferences.

All these interpretations attribute to individuals a coherent way of acting towards their own interests. In this case, economists did not have to bother about having a normative criterion at hand apart from preference-satisfaction. Indeed, albeit these interpretations provided different reasons that preference-satisfaction mattered, substantive ethical questions about *why* preference-satisfaction mattered did not to be asked under the assumption of coherent preference. As McQuillin and Sugden (2012) put it, 'economists could use a common theoretical system in which preference-satisfaction was the normative standard, while disagreeing about *why* preference-satisfaction mattered' (p. 555 — their emphasis). But once we seriously take into account that preferences are incoherent, interpretations about why preference-satisfaction matters — either in terms of *happiness*, *well-being* or *freedom* — suggest very different answers to the reconciliation problem.

Under the empirical evidence that preferences are incoherent, it seems that we cannot seriously hold that individuals seek to maximise their happiness, simply because we may find out that individuals do not choose according to what makes them happier. Instead, we may have to propose alternative measures of happiness that are not related to observed choice but to *hedonic experience*. The well-being interpretation faces the same faith. If preferences are incoherent, well-being cannot be grounded on the standard framework of observed preferences anymore since the large amount of cognitive biases documented in the literature suggests that individuals may not always act according to their own interests. Instead, we may have to attribute some preferences 'special' properties so that they can provide a better measure of one's well-being — typically the property that they are *undistorted from cognitive biases*. The freedom interpretation perhaps provides the most straightforward answer to the reconciliation problem. If preferences are incoherent, we may simply have to disentangle the idea that it is good to satisfy individuals' preferences because it is *their* preferences (the consumer sovereignty principle) from the preference-satisfaction concept. In other words, rather than assuming that the consumer sovereignty principle depends on coherent preference, we may instead focus on the institutional process that allows individuals to *enhance their opportunity to choose from*, disregarding whether their preferences are coherent or not.

By breaking with the assumption of coherent preference, it is as if normative economics opened the Pandora box from where the evils of justifying whether preference-satisfaction matter unleashed.[4] The point is if normative economics needs to be consistent with behavioural economics, it also needs to ground normative criteria on something other than observed choice. Those normative criteria can be grounded on the three ethical loci previously mentioned (*happiness*, *well-being* and *freedom*), but nothing forbids to ground normative criteria on yet other ethical loci unfamiliar with the tradition of normative economics (to be discussed in Section 3.5). The aim of the next section is to review

[4]This metaphoric transcription of normative economics obviously only gives the reconciliation problem an epic twist but should not be taken at face value. Eminent contributors to welfare economics and social choice such as Harsanyi (1977) and Sen (1991) already aimed at giving preference-satisfaction a substantial content without waiting any reconciliation problem to be recognised as such. Also, there have been long standing concerns about individual preference satisfaction: whether it makes any sense as a definition of well-being and its practical usefulness. See e.g. Hausman, McPherson, and Satz (2016) for how problematic individual preference satisfaction is as a standard for well-being without seriously addressing the impact of behavioural economics. What can nonetheless be said is that the accumulation of empirical evidence that individuals' preferences are incoherent is now large enough to recognise the reconciliation problem as a topic of its own.

some of the important methodological and theoretical issues associated with the main normative criteria offered in response to the observation that individuals have incoherent preferences. These normative criteria are *experienced utility* (happiness interpretation), *true preference* (well-being interpretation), *choice-basis* (well-being interpretation) and *opportunity* (freedom interpretation). The collection of those methodological and theoretical issues is a necessary step for asking what it takes for a normative criterion to be a 'good' normative criterion — a task I engage in Section 3.3.

3.2 A Critical Review of the Main Normative Criteria to Solve the Reconciliation Problem

3.2.1 Experienced Utility

The *experienced utility* criterion takes the happiness interpretation of the reconciliation problem. To contextualise things from a history-of-economic-thought perspective, it — strictly speaking — aims at measuring individuals' hedonic states in the line of Bentham's (1780 [2007]) meaning of utility as pleasure/pain calculus and of Edgeworth's (1881) concretisation of a hedonimeter.[5]

The conceptual appeal of measuring individuals' hedonic states is grounded on the theoretical discrepancy between what individuals do (what the authors refer to as *decision utility*) and what they experience (what the authors refer to as *experienced utility*). Since what individuals do is subject to many cognitive biases, the idea is to take only what they experience in terms of pleasure and pain as the normative criterion for evaluating their states of affairs. The ethical premise of the experienced utility criterion can thus be expressed in the following line. *It is good to maximise individuals' experiences of pleasure (or to minimise individuals' experiences of pain)*. The methodological and theoretical problems of the experienced utility criterion are various, but I shall restrict myself to the ones that are perhaps the most concerning (see Chapter 2 for a detailed review).

First, it is often argued that hedonism, when formulated into the maximisation of experienced utility, is too narrow a criterion to capture what makes the good life. This point is well acknowledged by tenants of the experienced utility criterion, who argue that hedonic experience is a component of the good but that the good life is certainly not empty of hedonic evaluation.[6] This is however already problematic for the ethical and general requirements because (i) there is a wide range of dimensions of the good life that is neglected with this normative criterion and (ii) one may wonder whether it is the goal of public policy to promote pleasurable experiences and not indirect measures of happiness such as access to public transport, green spaces, good air quality, etc. — and

[5]The literature includes Kahneman and Snell (1990, 1992), Kahneman and Varey (1991), Varey and Kahneman (1992), Kahneman et al. (1993), Fredrickson and Kahneman (1993), Kahneman (1994, 1999, 2000, 2011 [Part V]), Redelmeier and Kahneman (1996), Kahneman, Wakker, and Sarin (1997), Schreiber and Kahneman (2000), Redelmeier, Katz, and Kahneman (2003), Kahneman et al. (2004), Kahneman and Sugden (2005), Kahneman and Krueger (2006), Kahneman and Thaler (2006) and Dolan and Kahneman (2008).

[6]See Varey and Kahneman (1992, p. 169), Kahneman (1994, p. 21), Kahneman, Wakker and Sarin (1997, p. 377) and Kahneman and Sugden (2005, p. 176) who make that point explicit.

let individuals free to pursue whatever they want.

Second, experienced utility resurrects an ancient evil of standard welfare economics: the psychological character of interpersonal comparisons of utilities. The theory of experienced utility measurement is constructed of several axioms (Kahneman, Wakker, and Sarin 1997; Kahneman 2000). One axiom strictly assumes ordinal comparisons between individuals' utilities of different outcomes (e.g. one individual experiencing the taste of an exotic fruit, the other experiencing a guiding tour in Louvre Museum). The same axiom also assumes cardinal comparisons between individuals' utilities of the same outcome (e.g. two individuals experiencing the taste of the same ice-cream). This may be concerning for economists who are reluctant to hold that there is something 'scientifically relevant' to interpersonal comparisons of utilities, especially when cardinality is *stricto sensu* assumed. The point is albeit it could make sense to consider experiences of *pain* to be commensurable, it is far from absurd to consider experiences of *pleasure* to belong to another system of perception that works differently. This is due to some psychological phenomena such as the 'adaptation effect': individuals adapt to their circumstances to the point that they do not perceive an increase of pleasure similarly than another individual who is differently endowed (e.g. a rich individual compared to a poor individual).[7]

Third, experienced utility measurement has been given fundamental reconsideration by Kahneman himself in a recent interview given to *Hareetz* newspaper.[8] The theoretical construction of experienced utility measurement is grounded on *moment utility*: what is experienced here and now. The major point made by Kahneman (1999) is that 'policies that improve the frequencies of good experiences and reduce the incidences of bad ones should be pursued even if people do not describe themselves as happier or more satisfied' (p. 15). In other words, the author argued that only the maximisation of moment utility is normatively relevant, *even if individuals actually have a more pleasant memory of an experience with lower moment utilities*. However, nothing really says why moment utility should be given more importance than *remembered utility* (the global retrospective evaluation of a past experience). In fact, Kahneman recently reconsidered that hedonic measurement may not be what matters to individuals' objective happiness. As the author puts it:

> 'People don't want to be happy the way I've defined the term — what I experience here and now. In my view, it's much more important for them to be satisfied, to experience life satisfaction, from the perspective of "what I remember", of the story they tell about their lives. I furthered the development of tools for understanding and advancing an asset that I think is important but most people aren't interested in.' (Kahneman — interviewed by Amir Mandel in 2018)

This yields to a considerable change in our understanding of how objective happiness should be measured. Perhaps it should not be measured in terms of internal states of mind (immediate pleasure) but rather in terms of what individuals remember of their experiences (which is a more durable state of mind than immediate pleasure), or even in terms of consideration about what makes the good life that is not directly related to pleasure (e.g. access to public transport, green spaces, quality of the air, etc.). The latter interpretation of happiness refers to an objective conception of goodness in which

[7] See Kahneman (1999) who provide an extensive psychological defence on how to palliate this issue.

[8] Full interview is available at https://www.haaretz.com/israel-news/.premium.MAGAZINE-why-nobel-prize-winner-daniel-kahneman-gave-up-on-happiness-1.6528513.

pleasure is only a by-product of happiness but does not *constitute* happiness (see Chapter 2).

3.2.2 True Preference

Unlike the experienced utility criterion, the true preference criterion does not take such a tangible account of well-being (i.e. happiness defined in terms of pain/pleasure calculus) but assumes a general psychological state in which individuals have the ability to meet their actual intentions/interests that are represented by their true/latent/laundered preferences. From the social planner's viewpoint, these preferences constitute individuals' 'normative' preferences (i.e. what they *should* prefer).[9]

True preferences (by opposition to other kinds of preferences which would be 'false', or 'mistaken') are defined as preferences that an individual would have had, had she not been disturbed by rational foibles/biases/errors/mistakes/anomalies/cognitive disturbances. The representation of actual choice as a combination of true preferences and errors allows the social planner to take only true preferences as normatively relevant. The social planner's goal is to identify these errors and then to reconstruct/recover the individuals' true preferences through various social mechanisms.[10]

One main advantage of this normative criterion is that it does not require to interpret well-being so narrowly as the experienced utility criterion. In this manner, it may capture different aspects of life that individuals can find valuable, and it may also be more generalisable to various choice situations. Indeed, with this approach it is (presumably) up to individuals to define what their own well-being is. The ethical premise of the true preference criterion can then be expressed as follows. *It is good to satisfy individuals' preferences that are undistorted by cognitive biases*. Several issues are however associated with this normative criterion.

First, contrary to the experienced utility criterion, which is psychologically well grounded (no doubt pain and pleasure are real psychophysical phenomena that can somehow be measured), the true preference criterion shares nothing of this sort. To the question of whether there is empirical evidence for the existence of true preference, we can straightforwardly short-cut that no empirical study has so far supported this claim, nor actually the contrary. The fact that true preference lacks of psychological explanation is the main concern of Infante, Lecouteux, and Sugden (2016a) who point out two fundamental problematic principles: (i) even in possession of full cognitive capacities the latent process of producing true preferences is left unexplained, and (ii) decision

[9]The literature includes Bleichrodt, Pinto, and Wakker (2001), Madrian and Shea (2001), Benartzi and Thaler (2002), Camerer et al. (2003), Thaler and Sunstein (2003, 2009), Thaler and Benartzi (2004), Beshears et al. (2008), Loewenstein and Haisley (2008), Dalton and Ghosal (2011), Rubinstein and Salant (2012), Pinto-Prades and Abellan-Perpiñan (2012), Halpern (2015), Thaler (2018) and Sunstein (2019).

[10]Note that before a large amount of behavioural economists had a great interest in the true preference criterion, some authors already gave considerable support for it. The concept of true preference follows the one of Harsanyi (1977, pp. 29-30) in his defence of utilitarianism. Fine (1995) aimed at distinguishing the two concepts of true preference and actual choice from a social choice perspective. From a philosophical perspective, some authors also already defended the satisfaction of self-interested 'informed', 'rational', or 'laundered' preferences as what constitutes goodness (Gauthier 1986 [Ch. 2]; Arneson 1990; Goodin 1992). See also Railton (2003): 'an individual's good consists in what he would want himself to want, or to pursue ... free from cognitive error or lapses of instrumental rationality' (p. 54).

theory has no competence to legitimise a single correct way of framing a choice problem, which is accessible to any individual (even if 'super-rational'). Perhaps the most important issue of the true preference criterion is its inability to provide a convincing account that individuals would be better off if they conform to the behaviour of the Econ (Berg and Gigerenzer 2010). Indeed, it appears that true preferences refer to nothing else than rational preferences, i.e. *coherent* preferences. The fact that the heuristics-and-biases *positive* program is presented as a radical departure from the Econ but that its *normative* program is (paradoxically?) built on it is commonly recognised in the critical literature.[11]

Second, if the assumption of true preference lacks of psychological explanation, tenants of the true preference criterion may need alternative ways to justify its application. One way for the true preference criterion to be operationalisable is to provide the social planner with some meta-criteria on what is judged to be a better outcome over another. This is to palliate the problem that the social planner cannot *know* what individuals' true preferences are.[12] Several meta-criteria have been proposed by tenants of the true preference criterion, but none of them seem to satisfy the general requirement because they restrict the scope of the true preference criterion to a narrow range of application. Those meta-criteria are the following.

- *Dominance.* When one alternative strictly dominates another either in terms of outcome or risk, it may be assumed that the former is better than the latter. For example, we may assume that individuals' true preference are to save the maximum amount they can (e.g. they prefer more money to less when they will be retired). Based on this assumption, the social planner could set the maximum amount as the default option of the 401(k) plan. This meta-criterion is proposed by Loewenstein and Haisley (2008). The issue is that dominance can only apply to some circumstances where more can unambiguously be compared to less (typically monetary outcomes). Furthermore, the 'more is better' maxim may not necessarily be consensual among individuals. For example, one may not necessarily prefer the travel trip bundle $\{France, Italy, England\}$ to $\{France, Italy\}$, simply because she may not like to visit England. The disliked alternative added to the bundle (here $England$) may play out negatively in the individual's personal evaluation.

- *Evidential view (or folk beliefs).* This meta-criterion holds the idea that the choice architecture (or framing) is legitimised when there are 'good' reasons to believe that the behaviour being encouraged will actually improve the well-being of individuals being influenced by the social planner. For example, under the assumption that eating healthy, not smoking and saving more is better, the choice architecture should be framed in a way that it will encourage individuals to eat healthy, not smoke and save more. Tenants of the true preference criterion who support this meta-criterion in fact mean something closely related to Hausman's (2012) 'evidential view'. The 'evidential view' states that preference-satisfaction does not constitute well-being but provides reliable information about well-being. Instead of having an ethical theory at hand, the idea is that folk beliefs about what constitutes goodness are enough to make sense of what makes individuals better off. The platitudinous character

[11] See Berg (2003, p. 431), Berg and Gigerenzer (2010, pp. 147-148), Hands (2014, p. 398), Whitman and Rizzo (2015), Lecouteux (2016) and Dold and Schubert (2018).

[12] Rizzo and Whitman (2009) call this problem the 'knowledge' problem and Rebonato (2012) the 'interpersonal intelligibility of preferences' problem.

of the 'evidential view' is fully recognised by Hausman (2012), who argues that 'platitudes concerning what is good for people still have content ... economists know enough about the things that make lives good or bad that they can make sense of the evidential view of the relationship between preference-satisfaction and welfare' (pp. 92-93). It is however disturbing to see that the author also holds elsewhere that 'economists who believe that they promote well-being by satisfying purified preferences need to know what people's purified preferences are, not what they should be' (Hausman 2016, p. 28). The issue is that folk beliefs only allow to say what individuals' preferences *should be*, not what they *actually are*. Strictly speaking, characterising such meta-criterion as 'evidential' seems misleading: what kind of 'evidence' folk beliefs provide about what makes individuals better off?

- *Self-officiating (or 'as judged by themselves')*. We are then left with what individuals would express what they judge to be their own good. This meta-criterion is given the name of 'self-officiating' by Loewenstein and Haisley (2008) and 'as judged by themselves' by Thaler and Sunstein (2009). It states that on the condition that individuals clearly self-express their willingness to lose weight, stop smoking, stop procrastinating, etc., the true preference criterion applies.[13] For example, if overweighted individuals consistently state that they would be better off if they were slim, and if they deliberately state that a paternalistic policy would make them better off, then such policy would be ethically justified (Loewenstein and Haisley 2008). Leaving any philosophical consideration apart (one may fairly question which of the many individuals' preferences over time has/have moral authority over the others [see Chapter 5]), one may argue that economists or social planners specifically want to have a normative criterion at hand when those *ex-post* feedbacks are unavailable (Bleichrodt, Pinto, and Wakker 2001).

- *Clearly negative outcomes.* This meta-criterion states that when everyone would agree under 'common sense' that one outcome is clearly at the cost of individuals' interest, it appears relatively unambiguous that the true preference criterion applies (Loewenstein and Haisley 2008). Such meta-criterion is basically the 'evidential view' pushed at its extreme. Addiction, bankruptcy or paying the exact same product more expansively that it could have been afforded (ethical considerations such as fair trade or environmental protection left apart) are examples of clearly negative outcomes. Among all the meta-criteria here listed, this one has perhaps the most intuitive appeal: some things are just bad for everyone. Perhaps a tiny fraction of people would argue the contrary, but an overwhelming majority seems enough to make the true preference criterion operationalisable under the 'clearly negative outcomes' meta-criterion.

Considering these methodological restrictions on the applicability of the true preference criterion, it follows that the true preference criterion only makes sense in situations where distortions from rationality uncontroversially make individuals worse off. To take yet another example, consider a case where individuals would have to exploit their cognitive abilities more carefully (e.g. studying and comparing prices) if some markets such as water supply, food or phone contracts were not regulated (Heidhues, Johnen, and Köszegi 2020). This could eventually lead individuals to choose a complex tariff that

[13] See Reisch and Sunstein (2016), Reisch, Sunstein, and Gwozdz (2017) and Sunstein, Reisch, and Kaiser (2019) for empirical surveys about Europeans' acceptance of *nudges*.

does not minimise their costs. But it would be relatively uncontroversial to state that for a given bundle of alternatives (such as a phone contracts), individuals who have the choice between complex tariffs simply want to choose the one that minimises their costs. The point is that in some specific cases, a good choice is indeed (and uncontroversially) a choice undistorted by cognitive biases. This approach is given the name of 'preference regularisation' instead of 'preference purification' by Infante, Lecouteux and Sugden (2016a, pp. 18-21), who argue that the true preference criterion only makes sense in such uncontroversial cases. Perhaps the main problem of the true preference criterion, which will however always lurk in the background, is what Infante, Lecouteux, and Sugden (2016a) characterise as the difficulty of framing a choice problem in order to elicit one's true preference (see also Section 1.4 in Chapter 1). The point is even if we propose a theoretical framework for identifying mistakes (Köszegi and Rabin 2007, 2008; Bernheim 2016), decision theory has (again) no competence to legitimise a single correct way of framing a choice problem, which is accessible to any individual (even if 'super-rational').

Third, the oxymoron character of libertarian paternalism (Thaler and Sunstein 2003, 2009) — which is the dominant approach that takes true preference as its normative criterion — is subject to many philosophical problems.[14] From a philosophical viewpoint, when the social planner exploits individuals' cognitive biases to help them taking the best decision, it is merely impossible not to violate individuals' liberal principles. This is true even if individuals explicitly agree with the paternalistic character of the policy intervention. In other words, libertarian paternalism inevitably requires to make trade-offs between well-being and liberty/freedom/autonomy. But one can hardly increase the former while leaving the latter unchanged.

3.2.3 Choice-Basis

Like the true preference criterion, the *choice-based* criterion also takes the well-being interpretation, although it could also be associated with the freedom interpretation. This normative criterion can be seen as a subtle version of the true preference criterion since it suggests a compromise between the problem that actual choice diverges from well-being and the possibility to nonetheless keep *choice* as a satisfactory proxy of well-being (thus 'rescuing' somehow the consumer sovereignty principle). The choice-based criterion takes a step farther in not defining what makes individuals better off because it only considers a minimal psychological state of observation, attention, memory, forecasting or learning processes for normative assessments (thus leaving any ambiguity of the individuals' reasons for choice apart).[15]

The conceptual appeal of this normative criterion is shared by some economists who are reluctant to assess individuals' mental states for either epistemic or practical reasons.[16]

[14] See Mitchell (2005), Rizzo and Whitman (2009, 2019), Welch and Hausman (2010), Grüne-Yanoff (2012), Rebonato (2012, 2014), Hédoin (2015, 2017), Sugden (2017b) and Scoccia (2019).

[15] The literature includes Bernheim and Rangel (2004), Köszegi and Rabin (2007, 2008), Bernheim and Rangel (2007, 2008, 2009), Salant and Rubinstein (2008), Loewenstein and Ubel (2008), Bernheim (2009), Dalton and Ghosal (2012), Rubinstein and Salant (2012), Masatlioglu, Nakajima, and Ozbay (2012), Manzini and Mariotti (2014) and Bernheim (2016).

[16] See Bernheim and Rangel (2008, p. 156), Manzini and Mariotti (2014, pp. 343-344) and Bernheim (2016, pp. 24-25) who advance the argument that welfare economists should evaluate individuals' states of affairs according to individuals' own conception of goodness (not happiness nor true preference). They

It is indeed worth to emphasise that economists usually make of *choice*, and not subjective well-being reports, their privileged data. In this manner, they conform to the liberal tradition of standard welfare economics by making choice (or observed preference) the main normative criterion for well-being.

The ethical premise of the choice-based criterion can then be formulated as follows. *It is good that individuals choose undistorted from cognitive biases.* As we can see, the ethical premise of the choice-based criterion is identical with the ethical premise of the true preference criterion. The subtlety is that the privileged data is here not *preference* but *choice*. That is, the social planner is not required to have individuals' expression of preference but only to identity cognitive anomalies with the help of theoretical models that rigorously define what a mistake is (Köszegi and Rabin 2007, 2008; Bernheim 2016). The goal of the social planner is then to only take observed choice undistorted from cognitive biases as the normative-relevant data. The main issues associated with this normative criterion are nonetheless the following.

First, if the choice-based criterion is grounded on the same assumption than the true preference criterion (i.e. that what makes individuals better off is some psychological states free from cognitive biases), then how does it fundamentally differ with the true preference criterion? The simple answer is that (fundamentally), there is no difference between these two normative criteria. The choice-based criterion therefore meets the same critique of the true preference criterion: it only accounts for situations in which distortions from rationality make individuals worse off.

Second, although tenants of the choice-based criterion are reluctant to assess individuals' states of affairs by measuring individuals' level of happiness, they still make room for mental states by giving it an 'auxiliary role'. According to Bernheim and Rangel (2008), 'ancillary conditions [or frames] must be observable in principle; otherwise, we would not be aware that a choice anomaly (i.e., the dependence of choice on the ancillary condition) exists in the first place' (p. 162). Also, according to Manzini and Mariotti (2014), 'choice data alone may not be enough ... we do not dismiss as irrelevant data different from choices, such as verbal reports or direct information on the cognitive processes of decision makers. We argue that such data may be useful in an "auxiliary" role: they help the observer to make educated guesses about the reasons for the agent's choice, reasons that may be welfare-relevant' (p. 344). Ultimately, the choice-based criterion seems to encounter a disturbing paradox that is well emphasised by Dhami (2016):

> 'Choice-based models must address the issue of choices that depart from those expected under the rational benchmark. In a leading model, one deals with this issue by trimming-away the anomalous choices. However, such trimming-away necessitates the use of either non-choice data, or the invocation of a welfare criteria for trimming the choices, which is what one is trying to construct in the first place.' (p. 1577)

The issue is if the choice-based criterion ultimately depends on either experienced utility or true preference, what is (fundamentally speaking) the added value of this normative criterion?

Third, perhaps the most concerning issue of the choice-based criterion is that it is (presumably) not ethically grounded at all. Unlike the other normative criteria here

also argue that choice is a far less obscure and more available data than anything else.

reviewed (*experienced utility*, *true preference* and *opportunity*), only tenants of the choice-based criterion are reluctant to say anything on the ethical content of the normative-relevant domain. This is made explicit by Bernheim (2016), who holds the usual 'ethically neutral' stance of standard welfare economics. In his words, 'The conventional economic framework seeks to assess well-being without factoring in ... moral considerations, concerning which economists have no special expertise. I follow that tradition' (p. 18). But since normative criteria are, by definition, rules that tell us whether one outcome is better than another, there is merely no way of avoiding ethical judgements about what makes one outcome actually better than another. The question is, how can a normative criterion be 'normative' at all if it does not presuppose what makes one outcome better than another? Yet claiming that individuals are better off being undistorted from cognitive biases is already an ethical judgement regarding what constitutes goodness.

3.2.4 Opportunity

The *opportunity* criterion breaks with rational choice as the normative benchmark, which is a common point shared by the true preference and choice-based criteria. Recall that the true preference and choice-based criteria lean on a separation between rational reasoning and cognitive biases. By emphasising that incoherent preferences are not incompatible with normative analysis, Sugden (2004, 2018a) proposes a normative criterion of *opportunity*, according to which more opportunity for individuals is better than less, *independently of what their preferences are*.[17]

The aim of the author is to maintain the liberal tradition of economics against libertarian paternalism, which purpose is to combine liberal and paternalistic principles (yet unsuccessfully, as previously mentioned). Sugden's (2018a) two central criticisms is that there is no reason to assume that true preferences exist beneath the psychology of actual mental processing, and that the social planner's viewpoint is irrelevant because citizens above all are ultimately concerned about policymaking. The author ambitions to replace what he calls the process of 'preference purification' with the concept of 'opportunity for choice', which the latter focuses on enhancing individual freedom to choose. In this manner, Sugden's approach takes the freedom interpretation of the reconciliation problem.

The benefits of the opportunity criterion are twofold: (i) it avoids the problematic aspects of the true preference and choice-based criteria of determining what a decision 'free from cognitive biases' is, and (ii) it avoids the need of saying what constitutes goodness by instead leaving individuals be the best judge of their own good. The ethical premise of the opportunity criterion can then be formulated as follows. *It is good to promote individuals with more opportunities to choose rather than less*. Like the rest of the normative criteria previously reviewed, the opportunity criterion is however not unproblematic from both methodological and theoretical perspectives.

First, the opportunity criterion forbids to make comparisons between sets that are not nested. To give a simple illustration, consider the opportunity set $O_1 = \{x, y, z\}$ compared to the opportunity set $O_2 = \{x, y\}$. Here O_1 dominates O_2 under the opportunity criterion because O_1 contains all alternatives in O_2 (that is, x and y) *plus* an alternative that is unavailable in O_2 (that is, z). But what if we have one alternative in one opportunity set that

[17] The literature includes Sugden (2003, 2004, 2007, 2008, 2010, 2017a, 2018a).

is not contained in another, e.g. $O'_1 = \{x, y, z\}$ and $O'_2 = \{w, x\}$? Because Sugden (2018a) does not suggest that the *nature* of some alternative may provide more opportunities than others, the opportunity criterion is speechless on evaluating opportunity sets that are not nested.[18] The same issue applies for any other combination where one alternative is not contained in another opportunity set. Consider for example an extreme case where $O''_1 = \{r, s, t, u, v, x, y, z\}$ and $O''_2 = \{w\}$. In this case, we can still not say anything on whether O''_1 or O''_2 provides more opportunity, even if the cardinal of alternatives in O''_1 is by far larger than the singleton in O''_2. This point constitutes a challenge for the general requirement, as there are many situations left apart where non-nested sets can simply not be evaluated with the opportunity criterion.

Second, there exist the psychological phenomena of *choice overload* and *self-constraint*, which may challenge the ethical premise that more choice (or opportunity) is always better than less.

- *Choice overload*. This psychological phenomenon is identified as the feeling of being worse off by having too many alternatives to choose from.[19] Choice overload is popularised by Schwartz (2004 [2016]), who identifies the following negative feelings associated with it:
  - *Paralysis (or inefficiency)*. More alternatives create paralysis (i.e. it is difficult to choose something at all). A related psychological phenomenon is emphasised by Benartzi and Thaler (2002), who show that more opportunities lead to more complexity and then to an inefficiency in picking the best outcome.
  - *Decreasing of satisfaction*. Even if individuals are not paralysed, they may end up being less satisfied than with fewer options. The potential reasons are the following.
    * *Regret and anticipation regret*. An individual who faces too many alternatives would easily imagine what it would have been if she had chosen another alternative. This tends to increase the risk of regretting the chosen alternative.
    * *Opportunity cost*. This refers to the previous reason formulated in economical terms. If the opportunity set is large, it is easy to think about missing an opportunity, thus making the individual less satisfied with the chosen alternative.
    * *Escalation of expectations*. The more choice the individual has, the more demanding she may become. In other words, her expectations may increase with the increasing of available alternatives. This eventually makes her less satisfied than she would have been if she had the choice between fewer alternatives.
    * *Self-blame*. The opportunity criterion is grounded on the consumer sovereignty principle, according to which not only individuals are the best judge of their own well-being but are also fully *responsible* for their own

[18] This point actually refers to a complex debate in social choice about how to measure opportunity, and whether opportunity is measurable at all. I come back to this point below.

[19] The empirical literature includes Iyengar and Lepper (2000), Hutchinson (2005), Botti and Iyengar (2006) and Scheibehenne (2008). See also Iyengar (2010) for a nuanced overview with consideration of cultural backgrounds.

choice (Sugden 2004, p. 1018). Consequently, it becomes easier to blame oneself for not having made the 'right' choice.

Albeit choice overload is documented in many experimental studies, the meta-analysis of Scheibehenne, Greifeneder, and Todd (2010) concludes that the recognition of this psychological phenomenon is however unclear:

> 'The meta-analysis further confirmed that "more choice is better" with regard to consumption quantity and *if decision makers had well-defined preferences prior to choice*. ... To understand the effect that assortment size can have on choice, it will be essential to consider *the interaction between the broader context of the structure of assortments* — beyond the mere number of options available — and *the decision processes that people adopt*.' (p. 421 — my emphasis)

In other words, choice overload is not a proposition to be answered by a true/false dichotomy. It requires to be sensitive about all the explanatory variables that may either facilitate choice overload or not. Yet another meta-analysis of Chernev, Böckenholt, and Goodman (2015) identifies four factors that do facilitate choice overload: *choice set complexity*, *decision task difficulty*, *preference uncertainty* and *decision goal*. Unlike the empirical literature on incoherent preference, the empirical literature on choice overload provides far less homogeneous conclusions to safely advance that such psychological phenomenon can be considered as a stylised fact. Nonetheless, it seems not absurd to consider the large amount of empirical evidence which gives substantial support for choice overload to seriously consider it as an important aspect of economic reality.

One concerning problem of the opportunity criterion is that it assumes that individuals have well-defined preferences *prior* to choice. In this case, and as the conclusion of Scheibehenne, Greifeneder, and Todd (2010) state, providing individuals with more choice (or opportunity) is not problematic. But one may easily consider that individuals do not necessarily have well-defined preferences prior to choice, and I believe Sugden (2004, 2018a) is not against this idea.

One may also argue that the ethical premise 'more is better' highly depends on the *nature* of the alternatives. As Schwartz (2004 [2016], pp. 24-25) puts it, some alternatives are perhaps worth being available in large varieties (e.g. food at the supermarket), while other may not (e.g. public utilities, education or health insurances). There is indeed no *a priori* reason to assume that all the available alternatives in the economy are not perceived differently among individuals (i.e. either 'less-opportunity wanted' or 'more-opportunity wanted'). The bottom line is that there may be some alternatives that individuals would like to have more opportunity to choose from, but not other.

- *Self-constraint*. This psychological phenomenon is characterised as the explicit want to have less alternatives rather than more.[20] Unlike choice overload, self-constraint is something explicitly wanted by the individual. Therefore, it (theoretically speaking)

[20]The philosophical literature includes Elster (1979 [1998], 1983 [2016], 2000). See also Thaler (1980), who discusses situations where individuals voluntarily restrict their choices, deliberately not choosing so as to avoid psychic costs that the choices might induce.

perhaps constitutes a bigger challenge to the opportunity criterion, which (again) gives fundamental importance to individual responsibility (i.e. being the master of one's own choice).

To illustrate how self-constraint may challenge the opportunity criterion, consider the following case where an individual has two possible consumption alternatives *fruit* and *cake* that she can consume in periods 1 and 2 (Sugden 2018a, p 150).[21] The individual can choose between a fruit and a cake in both periods, so her opportunity set is defined as $O = \{\{fruit, cake\}, \{fruit, cake\}\}$. Now assume that same individual would like to constrain her opportunity set to only *fruit* in period 2 (for some reasons that are not the business of the social planner nor anyone else). She can choose between a fruit and a cake in period 1 but only a fruit in period 2. Hence, her opportunity set is defined as $O' = \{\{fruit, cake\}, \{fruit\}\}$. According to Sugden's (2018a) individual opportunity criterion, 'any expansion of a person's opportunity set promotes her interests' (p. 99). Therefore, O dominates O'. However, if we ground normative assessment on the consumer sovereignty principle, according to which we must give fundamental importance to the individual's choice because it is *her* choice, we must respect her will to restrict her freedom to choose and therefore rank her opportunity sets in a way that O' dominates O.

Paradoxically, the opportunity criterion suffers from a disturbing theoretical problem: it does not account for the interests of individuals who want to constrain their own alternatives without violating its principle of providing individuals with more choices rather than less. To put it simply, it does not account for individuals who would like to have *the choice of not having the choice*.[22] This limitation of the opportunity criterion is well recognised by Sugden (2018a). In his words,

> 'How far a regime of voluntary transactions should be regulated so as to support individuals in imposing constraints on themselves is a deep problem that generations of economists have struggled with. I can only say that my analysis, as I have so far developed it, abstracts from this problem.' (p. 151)

To this objection, we can argue (like Sugden does) that very little economic activities are concerned with self-constraint, so that self-constraint may not constitute a big challenge for the general requirement. But we may retort that individuals do not necessarily need to show an explicit want for self-constraint *before* their decision in order to argue that self-constraint is a serious limit to the opportunity criterion. Instead, the need for self-constraint may arise once individuals notice they would have been better off with less choice rather than more. In other words, it may be an *ex-post* rather than an *ex-ante* individual evaluation: individuals may simply not be aware of the value of self-constraint *before* figuring out that more choice can make them worse off.

[21] What follows is taken from Mitrouchev (2019, p. 143).

[22] One may object that as long as the individual has the choice between O and O', my critic does not apply because the individual is still free to choose whatever she wants. But if we take individual psychology seriously, we should account for the individual's possible conflicting preferences of each of her multiple selves (see Chapter 4). That is, we should account for the fact that she may like to avoid being tempted by the cake, so having the choice between O and O' would not help her because her 'morally responsible self' would like to have the choice of only O'. I however recognise that finding such 'morally responsible' self may be problematic from a philosophical viewpoint (see Chapter 5).

If we take the two psychological phenomena of *choice overload* and *self-constraint* together, one problem of the opportunity criterion seems to be that it is merely *a-psychological*. That is, it avoids giving any relevance to individuals' psychological processes. But recall that the initial issue of behavioural economists interested in normative analysis is specifically to propose normative criteria based on individual *psychology*. As for what is of the concern of the opportunity criterion, one may fairly ask whether the positive feeling of increasing choice is superior to the negative feeling of having more choice, all things considered. If the answer is positive we may have good reason to use the opportunity criterion for normative analysis. But if the answer is negative, we may have good reason to give more importance to psychological processes, feeling, affects, etc. (i.e. anything related to how individuals perceive their situation), rather than to provide individuals with more opportunity to choose *regardless what they would feel with more opportunity*. The disturbing stance endorsed by Sugden (2018a) is that even if individuals would actually be less happy with more opportunity, it would still justify enhancing individuals' opportunity to choose from. Yet many may find this principle a bit extreme: is not what individuals *feel/perceive* that ultimately matters?

Third, how to measure opportunity (and whether it is measurable at all) is a complex debate in social choice that is far from being consensual. At first sight, this criticism may sound unfair given the other normative criteria offered in the literature. Is measuring happiness consensual, given the many different interpretations 'happiness' can take; is the measurement of true preference (or choice free from cognitive biases) consensual when it requires to assume that true preference exists beneath the psychology of individuals? My point is, regardless the methodological issues associated with each normative criterion, happiness defined in terms of pain/pleasure calculus is well defined, so as the concept of true preference and choice free from cognitive biases. But is it the same for opportunity? Perhaps not, because the notion of opportunity is not clearly delimitated in social choice (Pattanaik and Xu 1990; Sen 1991; Sugden 2003, 1998, 2010). In a nutshell, there are at least three competing approaches offered in the literature.

- *Pure quantity*. Opportunity can simply be measured in terms of the number of alternatives contained in the opportunity (or choice) set. For example, to solve the problem of the opportunity criterion that non-nested sets are not comparable, we may simply say that $O_1'' = \{r, s, t, u, v, x, y, z\}$ provides more opportunity than $O_2'' = \{w\}$ because O_1'' contains more alternatives than O_2''. Obviously, the issue with the pure quantity approach is that it is merely naive: it exclusively counts the number of alternatives without distinguishing the *nature* of these alternatives (Pattanaik and Xu 1990). To palliate this issue, an alternative measure of opportunity could differentiate between the nature of the alternatives, which is perhaps a more convincing way to objectively define what it takes to have more opportunity. For example, it may sound relatively consensual that the opportunity set $O_1''' = \{blue\ car, red\ car, green\ car, yellow\ car, black\ car, white\ car\}$ provides less opportunity than the opportunity set $O_2''' = \{blue\ car, bicyle, train\}$, simply because the combination of alternatives in O_2''' are more diversified than the combination of alternatives in O_1'''.[23]

[23] Some may still argue that the definition of opportunity in terms of diversity yields to the philosophical problem that one cannot expect individuals to attribute the same value to the properties of available alternatives. For example, a rich individual who has a passion for car collections may give more opportunity

- *Potential preference*. Another measurement of opportunity is 'the range of preference that individuals might have had in relevant circumstances' (Sudgen 1998, p. 323). This approach is given support by Sen (1991), who argues that preference-satisfaction and freedom are very deeply interrelated. In this approach, opportunity-metric cannot be dissociated with what individuals would like to pursue because it is specifically in being able to satisfy their preferences that individuals have more opportunity. According to Sugden (2010), this measurement of opportunity is however problematic because it inevitably associates potential preference with a conception of what individuals reasonably/morally would like to choose. In other words, potential preference requires to define what goodness objectively is — an enterprise that liberal tenants of the choice-based and opportunity criteria aim to stay away from.[24]

- *Opportunity without metric (mutual advantage)*. Yet another approach to opportunity endorsed by Sugden (2010) is that opportunity cannot be measured because it would require to objectively define what it is (a stance that the author is opposed to). In Sugden's (2010) words, 'opportunity is an open-ended concept: often, we cannot specify in concrete terms what a person does or does not have the opportunity to do, or what the value is of the things that she might do' (p. 48). Although opportunity is not measurable in this approach, the point of the author is that we can say whether within a given economy all feasible opportunities have been made available — and this is what ultimately counts in the author's conception of opportunity. The problem of leaving opportunity without measurement is however that it may disappoint a lot of social choice theorists, who would be reluctant to say that there is no objective characteristic associated with opportunity, such as pure quantity or diversity.

Fourth (and along what has been discussed previously), the metaphysical interpretation of responsibility is interpreted as an axiom that one is required to accept. But one may easily argue that responsibility is not a characteristic that everyone is expected to have. The idea is that providing an individual with more opportunities is meaningless if such individual is not responsible for her own choice, nor autonomous enough to make her own decisions. Consider for example students who are offered a course list, and because of their inexperience and youth cannot be seriously held responsible for choosing among the many available alternatives (Schwartz 2004 [2016], p. 18). Are individuals 'responsible/autonomous enough' to benefit from more choice? Sugden (2004, 2018a) assumes so. But in situations where they cannot be expected to be, more opportunity may be harmful. Consider limited cognitive abilities of individuals facing complex and opaque information: one cannot always expect individuals to be perfectly informed about what they choose. The capacity of being able to make *enlightened* choices is then a serious concern for the opportunity criterion, where the concept of responsibility holds only if individuals are already well informed, well experienced, etc. Education plays here an

value to O_1''' than to O_2'''. To palliate this issue, opportunity in terms of diversity could be defended on the ground of public deliberation, i.e. that there are some alternative on which almost everyone would value the properties of available alternatives similarly.

[24]Yet this problem may not be that problematic for pragmatic purposes. For example, one could reasonably consider that there are some values such as *human capabilities* (Nussbaum 2000) that convincingly define what it takes to have more opportunity for every human being (e.g. being able to vote, to go to school, to have access to medical care, etc.).

essential role because it is not only a matter of having the 'right' type of information, but also a matter of *how* the information is conveyed. Yet this is an aspect neglected in Sugden (2004, 2018a).

Summary

To sum up, the table below resumes the methodological and theoretical issues associated with each normative criterion previously discussed.

Table 3.1: Summary of the methodological and theoretical issues of the *experienced utility*, *true preference*, *choice-based* and *opportunity* criteria

	Methodological and theoretical issues
Experienced utility	Hedonism is narrow dimension of what makes the good life
	Takes a peculiar conception of happiness in which only moment utility matters
	Assumes interpersonal comparisons of utilities (particularly cardinality)
True preference	There is no psychological support for the existence of true preference
	Can only be justified under specific cases where distortions from cognitive biases make individuals worse-off
	The dominant policy recommendation based on this normative criterion (libertarian paternalism) struggles to make paternalistic and liberal values compatible
Choice-basis	Based on the same concept than true preference, so the same critics apply
	Depends (paradoxically) on non-choice data
	Is (presumably) not ethically grounded at all
Opportunity	Forbids to make comparisons between non-nested sets
	Choice overload and self-constraint may challenge the idea that more choice (or opportunity) is always better than less
	Measuring opportunity is a complex debate in social choice
	Requires to endorse the axiom that individuals are responsible and autonomous beings under all circumstances

Because of these issues of different nature, the next section proposes a simple and unifying framework in order to assess the relevance of each normative criterion for normative analysis. The framework consists in three propositions that a good normative criterion should satisfy: the *general*, *ethical* and *practical* requirements. In the next section, I make these propositions explicit. I then present my result in Section 3.4 that no normative criterion satisfy all three requirements.

3.3 Three Important Requirements That a 'Good' Normative Criterion Should Satisfy[25]

3.3.1 The General Requirement

In order to explicitly define the general requirement, an intuitive idea would be to represent generalisability in terms of the cardinal of situations a given normative criterion can apply to. That is, assume N is the set of situations and $card(N)$ is its cardinal. A criterion R is generalisable if and only if $card(N)$ is 'sufficiently big'. The problem is that there is little hope we will ever reach a knowledge about what 'sufficiently big' means. Instead, we can be content with a negative definition of generalisability expressed in terms of what is 'not generalisable enough'. For example, we have seen in the previous section that the true preference criterion only applies in relatively uncontroversial cases where the *folk belief* or *self-officiating* criteria apply. Intuitively, this is enough to state that the true preference criterion does not meet the general requirement because its meta-criteria impose considerable restriction on its applicability. Likewise, the experienced utility criterion only applies to the evaluation of pain and pleasure, and the opportunity criterion only applies if opportunity sets are not nested. Thus, none of the normative criteria seem to satisfy the general requirement. We have then the following first proposition.

Proposition 1. *A good normative criterion is a criterion which does not restrict to particular situations but instead applies to a 'wide enough' range of situations.*

3.3.2 The Ethical Requirement

Next, we need to give importance to what the interests of individuals are. The argument is simple. If we consider that normative economics deals with a broad range of human motives, it cannot continue with *partial* representations of what makes individual better off but instead needs to account for *general* ethical principles of what living a fulfilling life means. In other words, it needs to entail the many different aspects of life that individuals can find valuable. The problem is obviously that every normative criterion captures a partial representation of what makes individuals better off. Recall that the experienced utility criterion implicitly takes hedonism as its underlying ethical theory, the true preference criterion takes preference-satisfaction and the opportunity criterion takes freedom. To put it differently, if we agree that a normative criterion necessarily maps (implicitly or explicitly) to one underlying ethical theory, and if one ethical theory has a different locus of what constitutes goodness than another, then a normative criterion can simply not entail everything that matters to individuals. Indeed, one may reasonably argue that normative criteria are proposed to evaluate *different* dimensions of life-fulfilment: happiness, well-being, freedom, etc., but not *all* of them.

[25] Awareness should be given that the following three requirements are presented as approximate global conditions rather than as formal conditions. The reason I use the term 'requirement' rather than 'condition' is specifically because I want to avoid the confusion between an *informal* and a *formal* condition. Speaking about a 'condition' implies to determine whether such condition is partially or completely satisfied (and if it is partially satisfied, to what extent). Since the three requirements are difficult to strictly delineate, and because it will take plain words rather than logic and mathematics to discuss whether the main normative criteria meet these requirements, I use the term 'requirement' as a synonym of 'informal condition'.

Although laborious, there is however good reason to continue looking for normative criteria that could entail what mostly matters to individuals. Indeed, it would be unfortunate to be restrained to a normative criterion that only accounts for a narrow dimension of the good such as hedonism. The ethical requirement is thus perhaps the most difficult to define for two reasons. First, holding that ethics should play out a significant role in normative economics is not mainstream. Second, even if one agrees that ethics should play out a significant role in normative economics, there is little hope that a convergence towards a single ethical theory will ever occur. To escape the second problem, we may need to draw a practical distinction between ethics and normative economics. For lack of consensus about what ethical theory is the 'right one', economists can go along with interpretations of what makes individuals better off without having to wait for philosophers to know what things are inherently good for human beings (if such thing can ever be known). This does not lead me to uphold Hausman's (2012) 'evidential view', according to which economists do not need an ethical theory in order to make normative assessments. My view is rather nuanced. Economists *do* need to uphold an ethical theory for justifying their normative approach, but this does not mean that they should consider that a given ethical locus *constitutes* the good. Instead, they can consider it to be a decent *representation* of the good.

Importantly, we also need to say that the ethical requirement is satisfied when the ethical judgement it proposes is shared among most individuals. For example, if most individuals think that goodness is defined in terms of happiness, and that happiness is defined in terms of social relationships, then 'social relationships' is a better ethical representation than some other component of happiness, say, pleasure. If they instead think goodness to be defined in terms of freedom and that they think freedom should be defined in terms of abilities to do what individuals want to achieve, then this is a better ethical representation than another — and so on. My point is that a normative criterion that entails what mostly matters to individuals should not be restrained to peculiar dimensions of what represents the good life. For this reason, a normative criterion should be ready to represent most interests of individuals. We have then the second proposition.

Proposition 2. *A good normative criterion is a criterion which is able to cut up between a desirable and a non-desirable state of affairs but which leaves the question of which theory of ethics is the 'right' one open because it uses ethical representations rather than ethical constitutions of goodness.*

3.3.3 The Practical Requirement

Lastly, we need a normative criterion that proposes a measure of what makes individuals better off. A decent measure would be a method that tells us the quantitative or qualitative level of the underlying value we are concerned with, e.g. happiness, well-being or freedom. Except the opportunity criterion, all of the normative criteria previously reviewed provide a measure of what makes individuals better off. Recall that the issue with the opportunity criterion is that the notion of opportunity is subject to debate in the literature of social choice. That is, another important point for the practical requirement is that the measurement provided should be relatively consensual among economists. The idea is to have a *universal* measure. Just like the metric system, individuals should first state a convention on which everyone agrees with, so that they can agree on what is actually to be measured. The third and last proposition is therefore expressed as follows.

Proposition 3. *A good normative criterion is a criterion which allows for measuring the values that are considered to be important for individuals, and which measure is 'relatively consensual'.*[26]

3.4 Assessing Each Normative Criterion with Respect to the General, Ethical and Practical Requirements

We are now ready to summarise in a table whether each normative criterion reviewed in Section 3.2 (*experienced utility*, *true preference*, *choice-basis* and *opportunity*) satisfy the *general*, *ethical* and *practical* requirements. In order to be more precise and to make the evaluation more informative, the *ethical requirement* can split in the following four separate questions and the *practical requirement* can split in the following two separate questions.

1. **General requirement.** *Can the normative criterion apply to a wide range of choice situations?* [theoretical and philosophical question]
2. **Ethical requirement**
    - 2.1. *Can the normative criterion capture the many different aspects of life that individuals can find valuable?* [ethical question]
    - 2.2. *Does the normative criterion actually capture the many different aspects of life that individuals can find valuable?* [ethical and empirical question]
    - 2.3. *Is the normative criterion supposed to be psychologically grounded?* [theoretical question]
    - 2.4. *Is the normative criterion actually psychologically grounded?* [empirical question]
3. **Practical requirement**
    - 3.1. *Does the normative criterion provide a measurement of individuals' states of affairs?* [theoretical question]
    - 3.2. *Is the measure relatively consensual?* [empirical question]

The table below summarises the evaluation of the four normative criteria which results from everything that has being said so far. The questions are formulated in a way in which a positive answer ('YES') corresponds to an advantage, and a negative answer ('NO') corresponds to a disadvantage.

[26] Note that the proposition is expressed in terms of *ethical* measurement rather than *well-being* measurement because a normative criterion does not necessarily relate to a measurement defined in terms of well-being.

Table 3.2: Summary of whether the *experienced utility*, *true preference*, *choice-based* and *opportunity* criteria fulfil the *general*, *ethical* and *practical* requirements

		Experienced utility	**True preference**	**Choice-basis**	**Opportunity**
1. General requirement	*1.1. Can it apply to a wide range of choice situations?*	NO (only experiences of pain and pleasure)	NO (only when distortions from rationality make individuals worse off)	NO (only when distortions from rationality make individuals worse off)	NO (only nested sets)
2. Ethical requirement	*2.1. Can it capture the many different aspects of life that individuals can find valuable?*	NO (hedonism: narrow dimension of the good)	YES	YES	YES
	2.2. Does it actually capture the many different aspects of life that individuals can find valuable?	-	NO (struggles to preserve autonomy)	NO (presumably not ethically grounded at all)	NO (choice overload and self-constraint)
	2.3. Is it supposed to be psychologically grounded?	YES	YES	YES	NO
	2.4. Is it actually psychologically grounded?	YES	NO (inner rational agent critique)	NO (inner rational agent critique + depends on non-choice data)	-
3. Practical requirement	*3.1. Does it provide a measurement of individuals' states of affairs?*	YES	YES	YES	NO
	3.2. Is the measure relatively consensual?	YES	YES	YES	NO (complex debate in social choice)

As we can see, some normative criteria fulfil more requirements than others, but it would be misleading to count the approach that generates the more 'YES' as the one that should be preferred because some issues may be more concerning than others. For example, the experienced utility criterion fulfils relatively well *2.3, 2.4, 3.1* and *3.2*, but poorly *1.1* and *2.1*, which are essential requirements for a good normative criterion. On the other hand, recall that one may say that self-constraint only plays a small part in economic life (and perhaps in the scope of what generally matters to individuals), so we could almost respond a 'YES' to *2.2* if we consider choice overload not to be something very much valuable to individuals (or significantly supported by empirical evidence).

3.5 Discussion

With the growing interest of academic research in the reconciliation problem, reconciling normative and behavioural economics involves interdisciplinary thinking at the intersection of theoretical economics, ethical considerations about the constitution (or representation) of the goodness, and practical applications of normative criteria. The central claim of this chapter is that solving the reconciliation problem requires to have a normative criterion that can apply to many choice situations, that can say something consistent about what makes individuals better off, and that allows for practically measuring their 'better off' states. In this survey, I propose a simple framework in order to evaluate whether the normative criteria offered in behavioural economics reasonably satisfy these three requirements. The result is that none of them satisfy them all. At their

current status, I acknowledge these requirements to be abstract representations that are not strictly delimitated, and that it is up to the reader to agree with the way I define these requirements. In response to this potential objection, I advance two points.

First, this literature review is a necessary step for nurturing the debate over the reconciliation problem, and is in this way addressed to all the authors who share great interest towards this hot topic of research. A critical review of each normative criterion like the one proposed in Section 3.2 (which obviously cannot be exhaustive) is certainly useful in order to have an overview about what the most concerning methodological and theoretical problems of each normative criterion are. More importantly, a literature review is useful in order to try giving the reconciliation problem a unified structure. Shall the consensus be about the idea that the reconciliation problem cannot be solved because no normative criterion satisfies all three requirements, it would already constitute a result (perhaps a disappointing one, but still a result).

Second, no single study can resolve methodological disagreements between leading experts in behavioural and normative economics about a complex issue, nor hope for a one-shot consensus among the community of researchers interested in the reconciliation problem. The point is, if some economists do not even consensually agree on whether normative analysis should be either preference-based or choice-based, I surely cannot expect economists to reach a consensus by anytime soon about how the reconciliation problem is best tackled. Yet this synthesising work has the merit of inviting economists to debate about whether the three requirements I here propose are relatively consensual. If they are, we can at least advance on the other laborious task of how to fulfil all three requirements.

Ultimately, the usefulness of the present work is that it can stimulate avenues of future research, mainly about proposing alternative normative criteria that are neither grounded on the *happiness*, nor the *well-being* nor the *freedom* interpretation of preference-satisfaction. Actually, there seems to be a major obstacle which drives many economists out of considering normative criteria other than preference-satisfaction as the sole measure of what makes individuals better off: their commitment to subjectivism as the 'right' underlying ethical theory. It is however important to remind ourselves that the three interpretations of *happiness*, *well-being* and *freedom* are only some among many ethical loci offered in ethics. Also, it is not because standard welfare economics is grounded on preference-satisfaction that we should necessarily continue in that direction.[27]

The point is albeit happiness, well-being and freedom are the three main interpretations of what makes the good life in normative economics, the vast field of ethics has obviously many other theories to offer about what makes the good life. Economists interested in normative analysis may be interested in those alternative ethical theories, just as they may be interested in happiness, well-being and freedom. I here conclude by briefly exposing two other ethical interpretations unfamiliar with the tradition of normative economics and which are, at their current status, still at an embryonic stage:

[27] Surprisingly, the strongest *status quo* bias ever observed in behavioural economics may not be in subject's behaviour, but (ironically) in economists' behaviour. See Davis (2016), who provides a sociology-of-science analysis about the practice of economists towards their own science. The author argues that economists often tend to experience strong loss aversion and *status quo* bias in the assumptions they make.

virtue ethics and *meaning*.

3.5.1 Virtue Ethics

An alternative normative approach that is worth being considered as a solution to the reconciliation problem is the *virtue ethics* approach proposed by Bath, Ogaki and Yaguchi (2015, 2017) and synthesised in Ogaki and Tanaka (2017 [Ch. 11]).[28] The authors start from the following tridimensional taxonomy of the 'relevant' ethical theories to normative economics: consequentialism (e.g. welfarism or utilitarianism), deontology (e.g. egalitarianism) and virtue ethics, which they note the latter to be largely ignored in normative economics. The authors argue that public policy is mainly a matter of meta-preferences over social states of affairs, such as preferences for e.g. less addiction and more patience. In their words, 'it is virtuous to be pleased about what we should be pleased about' (Ogaki and Tanaka 2017, p. 193). The authors consider meta-preferences to be explicit statements of what makes the good life, e.g. the ones implicitly endorsed by Thaler and Sunstein (2009) when based on the *folk belief* meta-criterion, such as 'smoking is bad' or 'eating healthy is good'.

The novelty of Bath, Ogaki and Yaguchi (2015, 2017) is to extend welfarism by adding virtue ethics as a component of the social evaluation. That is to say, they propose that social evaluation should not solely depend on a standard welfare function $W(u_1(x), ..., u_n(x))$, but also on what they call a 'moral evaluation function' $M(\psi_1(x), ..., \psi_n(x); \psi^*)$, where x is a social state, $u_i(x)$ the utility function of individual i, $\psi_i(x)$ a function that expresses properties of the endogenous utility function of individual i and ψ^* the ethical benchmark of the idealistic social good. For example, ψ^* can represent zero addiction, full patience or any other relevant ethical benchmark the society judges to be good based on a certain ideal of what it means to live a 'virtuous' life. We can then express the ethical premise of the *virtue ethics* criterion as follows. *It is good to satisfy individuals' meta-preferences over what is judged to be desirable for society*.

The virtue ethics criterion can then be interpreted as a variant of the true preference criterion, where (i) there exists an ethical benchmark of the idealistic social good ψ^* and (ii) the set of psychological states is not entirely a matter of *subjective* evaluation (what does the individual judge to be her own good?) but also a matter of *objective* evaluation (what does society judge to be individuals' good?). The 'social objective function' O is then a function of W and M. That is, $O(W(x), M(x))$ is a function that evaluates social states by considering both welfarism and virtue to be what constitutes goodness. The idea is that individuals can of course judge what their own good is, but there are also some things in life such as patience or being healthy that are 'ethically desirable' for everyone. *Who* judges what is ethically desirable for individuals? The social planner (again), as this is necessarily implied by the social objective function O. However, it is not excluded that the social planner is here only a representative of the individual members of the society who consensually agree on what the social good is.

From an ethical viewpoint, the major drawback of this approach is that it tries to combine two theories of ethics that seem hardly compatible: consequentialism (here

[28] See also Bruni and Sugden (2013) for a philosophical defence that the market economy can be seen as a sphere of virtue.

welfarism) and virtue ethics. The former is particular to the tradition of welfare economics and social choice. According to welfarism, the social good is determined by (and only by) the well-being of individuals (Sen 1977). However, virtue ethics is supposed to entail an informational basis that goes 'beyond' well-being, such as patience and no addiction (in the authors' own terms). Thus, we have, on the one hand, goodness defined in terms of well-being, and on the other hand, goodness defined in terms of acquired character-traits or dispositions that are judged to be good (e.g. patience or no addiction).

The question is, how are we to combine both of these ethical theories into a normative criterion? The approach of the authors seems a bit disappointing, as the diligent reader may have noticed that (i) they define $\psi_i(x)$ as a function that expresses properties of the endogenous *utility function* of individual i (that is, virtue is nothing more than a component that is already included in one's utility) and (ii) they do not say anything about when virtue can outweigh purely welfarist considerations (that is, trade-offs between preference-satisfaction and virtue remain unspecified). But was the goal not to introduce virtue ethics in order to *replace* welfarism, or at least in order to take an alternative ethical locus to individual and social evaluation *other than* happiness, well-being or freedom? Unfortunately, the virtue ethics criterion proposed by Bath, Ogaki and Yaguchi (2015, 2017) is ultimately a component of the generalised social welfare function O. That function can merely be seen as an extension of welfarism but not as an alternative that departs from welfarism.

3.5.2 Meaning

There is however another possible ethical locus for normative economics that is even less known in the literature: the *meaning* individuals give to the realisation of their states of affairs. This approach is given support in Loewenstein (1999), Karlsson, Loewenstein, and McCafferty (2004) and Dold and Stanton (2020). The aim of Loewenstein (1999) is to 'enrich' the notion of utility by incorporating the dimension of non-consumption into the utility function. The author particularly does so by taking the case of human activity that does not derive from pleasure. As Loewenstein (1999) argues,

> 'Despite the blossoming of the utility concept and expanding appreciation for the diverse determinants of utility, the list of human motives that have been codified in utility functions, and hence incorporated into economic analyses, remains seriously incomplete.' (p. 316)

Loewenstein (1999) takes mountaineering as an illustrative case, where individuals who undertake such activity do so for reasons that are not obvious from an external standpoint. Indeed, the experience of mountaineering is very often miserable. But the aim is specifically to account for those unilluminating reasons because they appear to have the most sense. If individuals undertake miserable activities such as mountaineering, learning the theory of music or writing a PhD thesis, it must be something other than happiness (or well-being, or freedom) that motivates them to choose or to experience something. Such approach could then entail the broadest range of what matters to individuals. In this matter, the *meaning* criterion could potentially better satisfy the general and ethical requirements than any other normative criterion here reviewed.

One reason that meaning would be of good use for normative analysis can be found in Davis (2011), who fairly argues that economic choices intimately relate to our human conditions. The argument is that the economic part of one's decision (whatever an

'economic choice' might mean) has become tremendously important in individuals' lives to the point that it is on choices regarding mortgage, where to live, saving plans, etc., that individuals plan their lives as a long-term process. Normative economics typically takes preference-satisfaction as its normative standard, but is speechless about what hides behind consumption and individual behaviour. Instead, a broader perspective normative economics could take is to consider every choice as meaningful in one's life. For example, individuals can buy a bungie dive ticket as a leisure activity, they can buy food at the supermarket, invest on financial markets, etc., and yet attribute a *meaning* behind all these decisions, *even if they do not derive any pleasure or satisfaction from consuming these goods or undertaking these activities*. The ethical premise of the meaning criterion would then be expressed as follows. *It is good to realise individuals' expectations about living a meaningful life*.

The particularity of meaning is that it belongs to a category other than happiness, well-being, freedom and virtue ethics. It is an ethical locus of its own kind that can be traced back in *existentialism*: a philosophical tradition unfamiliar with normative economics. To put it simply, what Bentham is at the experienced utility criterion, philosophers such as Sartre, Camus, Nietzsche or Kierkegaard could be at the meaning criterion. Further research could then aim to (i) argue more extensively why such normative criterion would matter as in Karlsson, Loewenstein, and McCafferty (2004) and Dold and Stanton (2020); (ii) formalise a model of preference formation in order to understand the psychological mechanism under which individuals choose according to what they believe is meaningful for them; (iii) argue that economic policy has good interest in focusing on enhancing the desirability of individuals to engage in meaningful choices as a source of life fulfilment.

Regarding the empirical applicability of the model, we could ask individuals to provide reasons before (expectations) and after (confirmation) their choice. Depending on their emotional affection towards their choices, we can classify some choices as 'more existential' than others, by still assuming that every choice is 'existential' but differs in degree. For example, a choice between two careers is 'existentially superior' than a choice between two brands of ketchup at the supermarket. The overall philosophy of this normative criterion would be to give importance to the meaning individuals want to give to their life, considering that there are obviously certain choices that are more meaningful than others. As a consequence, the aim of public policy could be to enhance meaningful choices or experiences in life, *even if those choices or experiences provide less happiness, well-being and freedom/autonomy*.

3.5.3 Concluding Remark

With all the normative criteria that have been discussed in the present chapter (*experienced utility*, *true preference*, *choice-basis*, *opportunity*, *virtue ethics* and *meaning*), I hope to have provided economists with a rich and concise literature review that will stimulate promising perspectives of research about the ambitious project of 'reconciling' normative and behavioural economics. I now hand it over to the reader.

CHAPTER 4

The 'View from *Manywhere*': Normative Economics with Context-Dependent Preferences

Abstract

In this chapter we propose a normative approach that accounts for individuals' context-dependent preferences. We emphasise that the problem of integrating preferences in the reconciliation problem (McQuillin and Sugden 2012) is akin to the problem of preference aggregation in social choice theory. We use this analogy to argue that unlike most of the approaches offered in the literature, economists should better focus on the process of preference integration: how individuals confront their conflicting views and form their normative preferences based on such conflicting views. After emphasising some theoretical issues of the welfarist and contractarian normative standpoints, we suggest an alternative standpoint which accounts for the process of preference integration: the process by which individuals' multiple selves start with conflicting preferences and end up with their own preferences. The originality of our approach is that instead of proposing a single acceptable context, which is either (i) exogenous to the normative representation of individuals ('view from nowhere'), or which (ii) only accounts for their behaviour but not their internal processes that lead to their own preferences ('view from somewhere'), we propose that normative economics can focus on how individuals may confront the views from different contexts so that they can form their own normative judgements. We call such normative approach the 'view from *manywhere*'.

Acknowledgements. Joint work with Guilhem Lecouteux (*Université Côte d'Azur*). We thank the audience of the 2019 workshop 'Behavioural Public Policies' in Nice, of the 2019 conference 'International Network for Economic Method' in Helsinki and of the REGARDS and GATE seminars for helpful comments. In particular, we thank Antoinette Baujard, Cyril Hédoin, Uskali Mäki and Kevin Techer for insightful discussions. All mistakes remain ours.

4.0 Introduction

Welfare economics traditionally assumes that individuals are defined by well-ordered preferences over the set of available alternatives — that is, preferences that are complete and integrated (non-stochastic, context-independent and internally consistent).[1] Individuals are supposed to behave as if they seek to satisfy their preferences and are deemed to be better off in a situation A compared to a situation B if and only if their preferences are better satisfied in A. Individual choices thus offer direct evidence for what improves individuals' well-being. Behavioural economics however documents many preference 'irregularities', highlighting that individuals' preferences are not well-ordered (Tversky and Thaler 1990; Kahneman and Tversky 2000; DellaVigna 2009). This raises serious concerns for normative economics because it is not certain that economists can take individuals' observed preferences as decent indicators of what makes them better off.

A common response to the challenge of building foundations for normative economics compatible with the findings of behavioural economics — challenge labelled as the 'reconciliation problem' (RP) by McQuillin and Sugden (2012) — is to draw an explicit distinction between the *behavioural* and the *normative* preferences of individuals. While *behavioural preferences* correspond to the preferences that guide individuals' actual behaviour and are inferred from their observed choices, *normative preferences* correspond to the choices individuals ought to do. Kahneman, Wakker, and Sarin (1997) for instance distinguish between 'decision utility' and 'experienced utility', and Beshears et al. (2008) between 'revealed preferences' and 'true preferences' (which they also label the latter as 'normative preferences').

Since it is not possible to directly infer individuals' normative preferences from their choices, most of the literature considers that it falls to the theoretician as an external observer to characterise those normative preferences.[2] When the theoretician does not imagine himself directly in such position, he usually addresses his recommendation to an abstract 'social planner' in the economist usual jargon (Kahneman, Wakker, and Sarin 1997), to a non-less abstract 'choice architect' (Thaler and Sunstein 2009) or to an actual consultant advising her clients (Harrison and Ross 2018). The standpoint endorsed to define what makes individuals better off is here *third personal*: it is the impersonal perspective of an external observer who thinks in a disinterested way about what is objectively good for individuals. Borrowing the term coined by Nagel (1986), Sugden (2013) calls such standpoint the 'view from nowhere'. One issue with this approach is that the theoretician relies on his *own* best judgements, i.e. it is assumed that individuals agree with the theoretician's assessment of the given situation.[3] For example, individuals are required to agree that it is 'obvious' that they prefer to be healthy (and therefore that they should limit their calorie intake), or to have more money when they will retire (and therefore that they should save more when they are young). But the theoretician

[1] By 'integrated' we adopt the terminology of Sugden (2018a). The internal consistency of preferences is usually defined by axioms such as transitivity or the sure-thing principle, which permit a utility representation of individuals' preferences.

[2] We use the generic term 'theoretician' to designate the economist, philosopher, ethical theorist, etc., who models the choice problem and who intends to offer a normative judgement.

[3] This is less true for an economist who is working as a consultant for a public or private organisation. The normative criterion to be used here typically depends on the nature of the contract between the economist and the organisation.

here merely takes for granted that individuals agree with his 'enlightened' judgement. Accepting unconditionally the theoretician's judgement and recommendation however poses a serious risk for individuals, whose opinions are not solicited. Writing about Sen's (1985) and Mill's (1859 [1972]) view on freedom and well-being, Sugden (2006) summarises the problem as follows.

> 'When political philosophy is written from the stance of the moral observer, the reality of these risks is too easily overlooked. In proposing his own conception of what is valuable, an author has to provide a reasoned defence of his position. In doing this, it is easy to slip into assuming that anyone who understands these reasons will find them convincing. Without noticing, we can make the transition from the belief that we are right to the belief that we will come out on the winning side of a reasoned discussion about what is right. So, we are inclined to think, we have nothing to fear from allowing evaluative issues to be resolved in a properly conducted democratic process. Indeed it is surprisingly easy to go further, and to imagine that the process has already been carried out, and everyone *has* agreed with us.' (Sugden 2006, p. 50 — his emphasis)

Sugden criticises Sen's approach for relying on such external judgement in order to define what individuals have reason to do (more specifically, the opinion of the majority). The author contrasts Sen's approach with the one of Mill, which instead allows individuals to satisfy '*whatever they might desire*' (Sugden 2006, p. 45 — his emphasis). Rather than defining the 'correct' preferences from an external standpoint, what matters in Mill's account is the freedom of individuals to achieve well-being according to *their own views* (i.e. not according to the one of an external observer). Here the adequate standpoint in order to judge individuals' normative preferences is then *first personal*.

The starting point of this chapter is the striking similarity between some questions raised by the RP (and more specifically the definition of an adequate normative criterion) and traditional debates in social choice theory.[4] The main solution offered so far to the RP is mostly *third personal* and committed to a welfarist approach when it comes to solving the problem of preference *integration* — just like most of the social choice literature engaged with the problem of preference aggregation. By adapting Arrow's impossibility theorem (1951 [2012]) to a multiple selves model, we show that the RP faces some major theoretical difficulties. Our alternative proposal is that the adequate standpoint to define individuals' normative preferences is neither third personal nor first personal but *second personal* (Darwall 2006). We adapt Sen's (2009) concept of 'positional views' and propose a normative criterion of 'individual awareness'. According to this normative criterion, the individual is judged to be better off or worse off depending on her degree of awareness about the factors that may influence her choice. The precise mechanism through which the individual weights those different factors in order to form her judgement is not determinate, as the theoretician is not legitimate to impose what he thinks would be reasonable to prefer *in fine*.

The rest of this chapter is organised as follows. Section 4.1 formally formulates the RP and highlights some similarities with social choice. Section 4.2 reviews the welfarist third-person standpoint — the 'view from nowhere' — and highlights its main shortcomings. Particularly, we adapt Arrow's impossibility theorem to the problem of

[4] This parallel is already explicit in Sugden's (2004) individual opportunity criterion. The author acknowledges that when thinking about the RP in the early 2000s '[he] followed the same broad strategy as [he] had done in responding to Sen's [impossibility of a Paretian Liberal] theorem'. Specifically, the author substitutes a criterion of opportunity to preference-satisfaction (Sugden 2018a, p. ix).

preference integration. Section 4.3 reviews Sugden's contractarian first-person standpoint — the 'view from somewhere' — and discusses some of its limitations. In particular, we emphasise that it difficultly deals with 'obviously' problematic choice situations such as drug addiction and self-acknowledged self-control failures. Section 4.4 develops our alternative standpoint — the '*view from manywhere*' — by emphasising the central place of the individual's evolving identity in the process of preference integration. Section 4.5 concludes by discussing some practical implications of our normative approach in terms of behavioural public policy.

4.1 Preference Integration and the Reconciliation Problem

One of the central and recurrent findings in behavioural economics is that individuals' preferences are context-dependent. That is, preferences may depend on seemingly 'irrelevant' aspects of the choice environment, such as the order in which the options are displayed or the way the choice problem is formulated (Tversky and Kahneman 1981). At the descriptive level, those inconsistencies revealed in individuals' behaviours can be accommodated by assuming the existence of *multiple selves* within the individual: each self being a traditional 'neoclassical' agent characterised by complete and integrated preferences (see Ross's (2014) 'neo-Samuelsonian' methodology). While this multiple selves model (MSM) offers an adequate representation of individual behaviour and conflicting preferences, it remains silent about the determination of the *normatively relevant self* on which the external observer could judge what preferences make individuals better off. The aim of this section is to set the RP as a problem of preference integration within the MSM, and then to use the tools of social choice to highlight the problem of preference integration. We first define what we mean by 'context', then introduce our two central notions of *behavioural* and *normative* preferences. Then, we formally formulate the RP as a MSM problem.

4.1.1 Defining the 'Context'

Suppose an individual I must choose an option x among the set of available alternatives X. This decision takes place within a specific 'context'. It is important to emphasise that while the notion of 'context' is routinely used in behavioural welfare economics (Bernheim and Rangel 2008) and seems rather intuitive, finding a precise definition of 'context' is a significant issue as it easily leads to circular definitions. We propose here a definition adapted from Larrouy and Lecouteux (2018) based on the premise that the context is what *we*, theoreticians, consider as the 'irrelevant' features of the choice environment when evaluating the choice problem of an individual (Bacharach 2006, p. 13).

We consider that each alternative $x \in X$ is characterised by a set of different properties P. For example, the moment and location in which the alternative x is chosen and the manner in which the information is presented are properties of x. We define a property as a function $P(j) : X \longmapsto \mathbb{N}$ that associates an index from the set of natural integers to each alternative.[5] When facing a choice situation, the individual is aware of a certain

[5]We do not need that properties map into the set of integers but more generally into a set that can be well-ordered by the property.

number of properties based on which she will compare the different alternatives. We denote $\mathcal{P}_I = \{P(j)\}_{j \in J_I}$ the awareness set of individual I. We must also consider that the individual's choice may be influenced by properties she is not aware of. We denote $\mathcal{P'}_I = \{P'(j)\}_{j \in J_I}$ the set of properties that influence the individual's choice without her being aware of it. This set typically includes the 'context' that we intend to define but could also include elements that could be welfare-relevant from the theoretician's own perspective. For example, the individual could strongly be convinced that all her actions are guided by pure altruism while she also tends to behave selfishly. For simplicity, assume the observer is in a higher epistemic position and is thus aware of both sets of properties. This means that as theoreticians, we know all the properties that determine the individual's choice (we will not consider cases of properties that are known to individuals but not to the theoretician). We can state this formally by defining $\mathcal{P}_E = \mathcal{P}_I \cup \mathcal{P}'_I$ as the awareness set of the theoretician.

The last ingredient we need in order to define the 'context' is the notion of 'relevant property' of the choice situation. We define $\mathcal{R}_E \subseteq \mathcal{P}$ as the set of properties that the observer considers as normatively relevant for the individual I, and the complement $\bar{\mathcal{R}_E}$ as the irrelevant features of the choice problem from the theoretician's own perspective. A context C is then a vector of values for the properties listed in $\bar{\mathcal{R}_E}$. We denote by $\Gamma \subset \mathbb{N}^{\bar{\mathcal{R}_E}}$ the set of possible contexts.[6]

4.1.2 Behavioural and Normative Preferences

We can now define our notions of *behavioural* and *normative* preferences. I's behavioural preferences when positioned in the context $C \in \Gamma$ is denoted $BP_C \subset X \times X$.[7] We interpret BP_C as a choice ranking (Hausman 2012). That is, '$x\ BP_C\ y$' means that I prefers x over y in context C. In other words, if I has to choose between the two alternatives when she is in the context C, she would pick x. There is no presupposition about the degree of consistency of BP (e.g. whether it is transitive) nor about its interpretation in terms of desires or motives for actions. It is just an analytical index aimed at representing the behaviour of the individual.

We now define $NP_C \subset X \times X$ as the normative preferences of the individual. While BP_C represents how the individual *does* behave, NP_C represents how the individual *ought to* behave. Crucially, we remain silent for the moment on the perspective from which we should define the 'ought to' principle (e.g. a transcendental normative principle, the opinion of the economist, the opinion of the individual, etc.). In particular, we discuss later whether $\mathcal{R}_E$ (the relevant properties of the alternatives according to the economist) should play a role in the definition of the individual's normative preferences. Our point for the moment is to represent the issues raised by the reconciliation problem in a com-

[6] We do not use the set of properties $\mathcal{R}_I$ that are considered as normatively relevant for I because it would require that I is aware of those properties — a cognitive situation akin to the 'inner rational agent' (Infante, Lecouteux, and Sugden 2016a). Moreover, it is not clear how the theoretician could solve the epistemic problem of recovering those properties. Furthermore, the whole challenge of normative economics (both standard and behavioural) is to define individuals' normative preferences with only $\mathcal{R}_E$ at the theoretician's disposal. That is, the theoretician does not know what could be the counterfactual $\mathcal{R}_I$ if individuals were in position to define their own well-being.

[7] Unless a confusion is possible, we note the preference relation without superscript designating the individual I.

mon framework, where the distinction between BP_C and NP_C allows us to distinguish between the positive and the normative dimensions of the individual's behaviour.

While we can observe the behavioural preferences of individuals — we indeed know the features of the various contexts $C \in \Gamma$ (which is a representation of the theoretician) and can therefore simply observe individuals' choices — we have more difficulty to define and observe their normative preferences. A natural constraint on the definition of NP (which is consistent with the common practice in welfare economics) is the principle of 'normative individualism' (Ross 2005, pp. 220-222). According to this principle, the proper locus of normative concern is individual persons, whose values and situations should be taken into account when debating ethical issues such as policy or justice. We translate this principle in our framework as follows.

Normative individualism (NI). $\forall C \in \Gamma, \exists C' \in \Gamma \mid NP_C = BP_{C'}$

This principle establishes a close relation between the behavioural and the normative preferences of the individual. It states that there exists at least one context C in which what the individual ought to do is simply what she would actually do in a counterfactual context C' (which may be the same context C or another one). The fundamental idea of this definition is that what makes the individual better off should not be set *a priori* but rather inferred from her actual choices, although possibly — but not necessarily — in a different context than the current one. Since NI is implicit in most works in normative economics we do not discuss it further. We now introduce two stronger assumptions.

Behavioural context-independence (BCI). $\forall C, C' \in \Gamma, BP_C = BP_{C'}$

Normative context-independence (NCI). $\forall C, C' \in \Gamma, NP_C = NP_{C'}$

BCI is an assumption that is subject to empirical test. It states that individuals' behaviours are not affected by the context in which they are embedded in when they choose. NCI is a normative claim about the proper definition of normative preferences. It states that the normative standard to assess individuals' well-being does not depend on the context in which they are embedded in. As behavioural economics questions the assumption of *behavioural* context-independence, it however remains less decisive regarding the assumption of *normative* context-independence.

Note that NI, BCI and NCI are consistent with the common practice in standard welfare economics, where it is assumed that the context does not play a significant role on individual behaviour. The usual argument is that the context may play a transitional role but should disappear when individuals adjust to the 'rational' pattern of behaviour (Harsanyi 1977) or under the 'discovered preferences' hypothesis (Plott 1996). Combining NI, BCI and NCI yields the following result.

Neoclassical consumer sovereignty (NCS). $\forall C \in \Gamma, NP_C = BP_C$

The result is an almost trivial corollary of NI since standard welfare economics does not consider the influence of the context on individual behaviour. Once we accept NI and acknowledge that the context plays no role, the principle of consumer sovereignty

follows directly. It is noteworthy that although welfare economists mostly agree with NCS, nothing is said about its ethical content. What matters is that individuals satisfy their preferences, although we do not need to agree on *why* it is a good thing that they satisfy their preferences. According to McQuillin and Sugden (2012), we may argue that satisfying preferences is desirable because it maximises one's happiness (Kahneman, Wakker, and Sarin 1997), because it maximises one's subjective well-being (Thaler and Sunstein 2003, 2009), or because it let individuals free to choose whatever they want to choose whenever they want to choose (Sugden 2004, 2018a) (see Chapter 3).

However, once we acknowledge that there may be a significant gap between individuals' behavioural and normative preferences, it becomes necessary to clarify the normative standpoint of the situation that needs to be assessed. This constitutes one of the essential challenges of the reconciliation problem.

4.1.3 The Reconciliation Problem

Behavioural economics provides extensive evidence that BCI is empirically inadequate: individuals' behaviours are affected by the context in which they are embedded in. To provide an illustration, consider the following choice problem of an Asian disease which is expected to kill 600 people (Tversky and Kahneman 1981, p. 453). The choice between the two alternative programs to fight the disease can be framed either as gains or losses (the number of subjects for each frame and the frequency of choices are specified in brackets).

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

Frame L [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%]

In the example above we can clearly see that behavioural preferences are likely to be context-dependent for many subjects (although some subjects consistently choose A & C or B & D). The issue for the theoretician is then to determine which program makes individuals 'better off' (if we can use this expression when facing such a stark choice), knowing that the preferences revealed through their choices are likely to depend on the way the problem is framed.

The RP could thus be stated as follows. Knowing that $BP_{C'} \neq BP_{C''}$ for two different contexts $C', C'' \in \Gamma$, how do we infer NP_C? The way we define NI permits some contexts to imply 'errors' on behalf of the individual. Specifically, what the individual chooses in certain contexts is not what makes her better off (e.g. choosing to drive after having too many drinks).[8] We however lack a decisive principle that could guide us to determine

[8]The most significant problem of driving under the influence of alcohol is that someone else may be hurt, although we only consider here the risk for the driver herself. That is, independently of whether she may cause an accident involving a third party, it is preferable from the perspective of her own well-being not to drive in such situation.

which context is the normatively relevant one. While this may be relatively obvious in the case of the drunk driver, the case of the Asian disease is a more concerning dilemma. Indeed, in cases such as the drunk driver the theoretician can rely on 'platitudes' about what makes the good life in order to know what is best for individuals (Hausman 2012, pp. 92-93). But in situations where negative outcomes are far from being clearly identified, platitudes about what makes the good life would typically be inadequate here.

Thus, the important question that must be solved when defining what the individual *ought to do* (her normative preferences) is *'according to whom?'*. While most behavioural economists consider that context-dependent preferences are problematic from a normative viewpoint, it is also because what counts as the context is the result of their *own* value judgements about what is 'relevant' in the given choice problem. For example, an individual may consider that the relative position of the fruit and the cake at the cafeteria counter is a relevant property of the choice problem. But it is unclear why the theoretician should be entitled to impose her own normative view about what counts as a relevant property.

Borrowing a terminology from metaethics, we distinguish three perspectives the theoretician can endorse to offer such ethical judgement: the *third-person*, *first-person* and *second-person* standpoint. The third-person standpoint takes an *outside* position and confers the duty of normative assessment to an external third party. It is the 'view from nowhere'. The first-person standpoint takes the perspective of the individual *herself* embedded in a current context. It is the 'view from somewhere'. The second-person standpoint, on the other hand, takes the standpoint of 'other' individuals — and more specifically of the multiple selves of the individual as a collective entity. It is the 'view from *manywhere*'.[9] While this last proposal is originally developed to take into consideration interpersonal ethical relationships, we here adapt it to consider *intrapersonal* ethical relationships. As an enduring person, I knows she can find herself embedded in various contexts (which may influence her behaviour) and there are good reasons to consider that — apart from cases of dissociative identity disorder — all the selves that constitute her person are worth being considered in the final ethical judgement of I.

In order to illustrate the difference between the third-person, first-person and second-person standpoints, consider the case of the drunk driver. If we assume there is no risk at all of involving someone else in an accident (otherwise all perspectives could reject the choice to drive based on this prospect), the first-person standpoint considers that it falls to the individual herself, when leaving the bar, to choose whether to drive or not. Taking the risk of being hurt in an accident or being fined is here a purely personal choice and should not be determined by a third party. This is, in a sense, akin to Mill's (1859 [1972]) 'Harm Principle'. According to the third person standpoint, there is a third party (the bartender, another client, a friend) who assesses the problem on behalf of the individual and advises her to drive or not. Note that in practice it is not certain at all that the third party in question is of good advice — although we assume here that when considering the recommendation of the theoretician, it is a *benevolent* advice. On the other hand, the second-person standpoint would imply that the individual should consider her other

[9]We thank Uskali Mäki for suggesting us this neologism rather than the term 'view from everywhere'. Indeed, 'everywhere' would mean that we intend to be exhaustive in the confrontation of all possible viewpoints of the multiple selves. We come back to this point in Section 4.4.

selves when choosing whether to drive or not (e.g. the one who may be put in jail later, the one with an injury and the one who miraculously came back home safe). The choice of not driving would here only be motivated by the harm she would cause to herself, put possibly in a different context (e.g. herself tomorrow morning with a serious headache).

As a complementary illustration, consider again the Asian disease experiment. According to the first-person standpoint there is not necessarily a problem in being inconsistent across frames. It indeed falls to the individual in each choice problem choosing the best program. The framing can also be seen as a guide to choose in such dilemma (Jullien 2016). The third-person standpoint — when the outside observer considers the two choice problems being identical — would suggest there is one 'correct' choice (e.g. either A & C or B & D) but determining such 'correct' choice would require introducing an additional external criterion to disentangle between the two options (see Section 1.4 in Chapter 1). On the other hand, the second-person standpoint would consider that the individual when embedded in a particular context (say, G) should also imagine herself in the other context (say, L). It is then only when the individual becomes aware of the two perspectives on the same problem (focusing on gains or losses) that she will be able to make an informed choice. Note however that she can still choose A and then D. What matters is not her *final* choice (nor its coherence) but that she is able to confront the different views prior to her decision, so as to avoid a 'pure' framing effect.

The question of which standpoint to endorse in order to choose the normatively relevant frame can thus be modelled with the tools of social choice theory. Here we consider that an individual is defined by several behavioural preferences $BP_1, BP_2, ..., BP_n$ corresponding to n different contexts. Recall that the context is a notion defined by the theoretician (or external observer). We make the additional assumption that each behavioural preference is itself complete and integrated for a given context. If the theoretician observes that this is not true for BP_C, he can refine his partition of the list of properties that define the alternative and make a finer partition allowing to rationalise a non-integrated BP_C as the interaction of finer contexts.[10]

Our primitive is therefore the coexistence of multiple selves who are individually 'neoclassical'. If I is in context C she will behave as described by the preferences of her self I_C. When placed in another context C' she will act as described by her self $I_{C'}$, and so on. I is however still the *same* (numerically identical) individual when embedded in those different contexts. We designate this continuing individual as I^*.[11] When looking at normative preferences, i.e. what makes I better off, we are interested in the normative preferences of the continuing individual I^*. Our objective is now to define the normative preferences of I^*, knowing that the only elements at our disposal (because we can observe them) are the behavioural preferences of the various I_C. In other words, the RP consists in *integrating* I's multiple behavioural preferences into the normative preferences of the continuing individual.

[10] Savage (1954 [1972], p. 88) shows that this is technically possibly if we can define an arbitrarily finite set of states of the world.

[11] Using the vocabulary of philosophy of identity, we here assume a unified view of the individual in which I^* is a persisting living entity connected by either psychological, physical, narrative or social relations. We however acknowledge this assumption not to be unproblematic from a philosophical viewpoint. See Chapter 5.

BP^* denotes the behavioural preferences of I^* (which are likely to be non-integrated), NP^* her normative preferences (which constitute our main subject of inquiry), while BP_C and NP_C denote the preferences of I^* when embedded in context C. Recall that under the principle of normative individualism, there exists at least one context C in which the normative preferences of I are defined by what she would do in a context C' (the same as C or not). We can now clearly see the analogy with social choice theory. While social choice theory starts from the preferences of the various individuals composing a society and investigates how to *aggregate* those preferences into a normative *social* function, we start from the preferences of the various selves composing an individual and investigate how to *integrate* those preferences into a normative *individual* function.

Starting from ordinal rankings BP_C to define NP^* (as suggested by NI) necessarily requires discussing whether Arrow's impossibility theorem can be meaningfully applied here. Following Steedman and Krause (1986), we argue that the four conditions of Arrow's impossibility theorem (Unrestricted domain, Pareto, Independence of irrelevant alternatives and non-Dictatorship) can be reinterpreted within a multiple selves framework. The result is that it may not be possible to define an overall ordering NP^* that is compatible with the multiple BP_C. In particular, we argue that the welfarist approaches in response to the RP necessarily violate at least one of these conditions.

4.2 Welfarist Approaches: The Third-Person Standpoint

4.2.1 Arrow's Theorem in a Multiple Selves Model

Recall that defining NP from the third-person standpoint implies to take the external standpoint of the theoretician in order to define what counts as 'well-being', and therefore the adequate procedure to integrate the preferences of the individual. Different normative approaches following this path are offered in the literature. However, each of those normative approaches violates at least one of the following conditions of Arrow's theorem.

- **Unrestricted domain (U).** We should not put any restrictions on the rankings BP_C of the various selves.
- **Pareto (P).** If $\forall C \in \Gamma, x \; BP_C \; y$ then $x \; NP \; y$.
- **Independence of irrelevant alternatives (I).** $\forall x, y \in X, \forall BP_C, BP'_C$ if $x \; BP_C \; y = x \; BP'_C \; y$ then $x \; NP_C \; y = x \; NP'_C \; y$.
- **non-Dictatorship (D).** The overall normative preference should not always follow the behavioural preferences expressed in one specific context. In other words, we should allow for $NP^* \neq BP_C$.

The theorem states that it does not exist a ranking NP that satisfies the four conditions (U), (P), (I) and (D) (Arrow 1951 [2012]). We can now realise that the four conditions have still some normative appeal in a MSM. (U) is indeed reasonable as it implies that the theoretician should not be restricted to analyse cases of 'reasonable' conflicts between preferences. Many cases of preference reversals, as in the Asian disease problem, however reveal strong conflicts. (P) is also a direct implication of NI. It states that if all the selves 'behaviourally' prefer an alternative x compared to y, it should also be the case for their

normative preferences. In other words, since normative preferences are defined from behavioural preferences it cannot be the case that an alternative x is preferred by all the selves, while not being at the same time normatively preferred. (I) sounds an *a priori* reasonable principle. Rejecting it would typically involve the types of preference reversals that is at the core of the RP. As for (D), it simply means that no normatively relevant self can impose her will and preferences over the others. Rejecting (D) would also require identifying which self is normatively relevant but it does not give any indication about how to select it.

We now review the different welfarist approaches of the RP and highlight which of the conditions (U), (P), (I) or (D) they violate.

4.2.2 Experienced Utility

The idea of this normative approach theorised by Kahneman, Wakker, and Sarin (1997) is to distinguish between 'decision utility' — which is the weight of an outcome in a decision — and 'experienced utility' — which is the hedonic quality of an experience in terms of happiness (see Chapter 2 for a detailed overview of this normative approach). Translated into our framework, decision utility refers to behavioural preference and experienced utility refers to normative preference. As presented in the Asian disease problem above, individuals' preferences are typically subject to change because of framing. Since decision utility is context-dependent BCI is here rejected. Also, since decision utility is inferred from observed choices and since observed choices are sometimes subject to cognitive biases, the idea is that individuals may not always choose the outcome that makes them better off. The rejection of NI comes from the fact that the normative criterion is unrelated to individual behaviour. That is, there is no obvious reason that there exists a context such that BP_C perfectly matches the maximisation of happiness.

The normative stance suggested by Kahneman (1999) is to define 'objective happiness' according to a set of normative rules that are external to the subject. The experienced utility approach therefore keeps NCI, where the locus of well-being is located in individuals' objective happiness. The rejection of BCI and NI, on the one hand, and the conservation of NCI, on the other hand, results in rejecting NCS — the principle according to which what individuals choose is what makes them better off in every possible context. Indeed, if no ranking of BP_C corresponds to 'happiness' then NI is rejected. But if a ranking BP_C corresponds to the ranking of 'happiness' then non-Dictatorship is violated. In this case, the experienced utility approach would impose the normative preference of one of the selves to the others. This however sounds arbitrary and may require further ethical justification.

4.2.3 True Preference

The true preference approach takes the satisfaction of individuals' preferences that are not distorted by cognitive biases as normatively relevant. Perhaps the most famous account is given by Thaler and Sunstein (2003, 2009) in their defence of libertarian paternalism. Just as in the experienced utility approach, the true preference approach starts from the observation that behavioural preferences change across context. This means that BCI is rejected. It also recognises that observed choices are sometimes subject to cognitive biases,

which means that individuals may not always choose the alternative that makes them better off. Yet the true preference approach does not reject NI because it assumes there exists an adequate context in which the behavioural preferences of individuals are equal to their normative preferences, 'as judged by themselves' (Thaler and Sunstein 2009). The adequate context in the definition of NP is when individuals have 'complete information, unlimited cognitive abilities and no lack of self-control' (Thaler and Sunstein 2003, p. 175). This abstract cognitive state is sometimes labelled as the one of the 'inner rational agent' (Infante, Lecouteux, and Sugden 2016a). Translated into our framework, NCI is therefore maintained. In this approach, NCS is however rejected because paternalistic interventions are often designed to help individuals take the best decision free from cognitive limitations.

With our framework, two objections can be made against the use of true preferences as the normative preferences. As argued by Infante, Lecouteux, and Sugden (2016a), assuming that the individual is free from cognitive limitations does not necessarily guarantee she will *in fine* generate preferences that will be complete and integrated. If we interpret the inner rational agent as a counterfactual entity (i.e. as what she would prefer if she was 'perfectly rational' and non-biased) we need an algorithm indicating how the individual integrates the various facets of her preferences over alternatives. The intuition that is informally stated by Infante, Lecouteux, and Sugden (2016a) can be formally proven if we interpret the problem of the preferences of the inner rational agent as a question of preference integration.

For the reasons discussed above, it is not clear which of the conditions (U), (P), (I) or (D) should be rejected in order to permit the definition of one's true preferences. Rejecting (U) would indeed restrict the analysis to 'not-too-problematic' cases; rejecting (P) is at odds with the 'as judged by themselves' clause; rejecting (I) could imply the type of preference reversals that are problematic in behavioural welfare economics. But if we keep (D) we know that we *cannot* define true preferences for the inner rational agent as complete and integrated.

A possibility would be to impose the preferences of one of the self of the individual (hence rejecting (D)) by suggesting that the inner rational agent is indeed one of the self of the individual. This means imposing *consistency* as a normative criterion: her 'true' self would thus be her neoclassical *alter ego*. Imposing such normative criterion is however controversial and would require some additional justification.[12] Finally, (U), (P), (I) and (D) are necessary conditions for the definition of an overall ordering but they do not provide the actual algorithm for integrating preferences. The only information the theoretician has is that the individual's preferences are complete and integrated. However, the theoretician cannot know *a priori* whether the individual is better off being risk-averse or risk-seeking in e.g. the Asian disease problem. In other words, the theoretician would merely know that the individual would choose consistently across the two problems but he would not know which program the individual ought to choose.

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

4.2.4 Choice-Basis

The aim of this approach advocated by Bernheim and Rangel (2008) is to extend standard choice welfare analysis to situations where individuals make 'anomalous' choices of various types commonly identified in behavioural economics. In this approach, frames are (by assumption) irrelevant to well-being. The scope of this approach is about the identification of welfare-relevant choices that are involved as means in reaching a certain outcome. The aim is to find an operational misunderstanding of the relationship between means and outcomes (such psychological process being labelled a 'mistake') that can be elicited with the use of cognitive data, precisely the lack of observation, attention, memory, forecasting and learning processes of individuals (Bernheim 2016). The social planner's goal is then to delete 'anomalous' (or 'suspect') choices in order to construct an individual welfare function.

The main difference with the experienced utility and true preference approaches is that the choice-based approach aims at preserving BCI. The choice-based approach does take for granted that individuals' preferences may change across contexts (just as in the Asian disease problem). But for the sake of normative analysis, BCI is 'rescued' as it is argued that choice remains the main guide for welfare analysis in the standard framework.[13] NCS is then deducted from the restricted set of choice data which is considered to be 'unbiased'. The argument is that NCS can be maintained if BCI is 'rescued' so that observed behaviour remains a good indicator of what makes individuals better off. Just like the true preference approach, NI and NCI are also preserved.

Within our framework, removing the ambiguous data from welfare analysis however violates the Unrestricted domain condition. We cannot indeed form normative judgements about cases in which the behavioural preferences of the individual are too 'conflictual'. This means that the range of situations which can be studied is rather limited in this normative approach. Again, ambiguous cases such as the Asian disease problem are then left apart.

4.2.5 Quantitative Intentional Stance

Another normative approach called the 'quantitative intentional stance' does not infer normative preferences from a given criterion but estimates them as the 'most plausible econometrically' (Harrison and Ross 2018; Alekseev et al. 2019).[14] This is made possible with a structural estimation of the underlying model of individual choice. Rank-dependent utility, together with expected utility theory, are assumed to provide decent normative guidance, although the former implies a non-linear treatment of probabilities. The authors claim that the adequate context to elicit normative preferences is the lab because it is a 'small world' where there is little room for a 'noisy' context that can influence individuals' preferences. In this approach, possible errors come from stochastic noises, which can be estimated econometrically.

[13] We understand the choice-based approach as a pragmatic consensus. It extends the revealed preference framework by taking into account the cognitive processes of individuals without modifying its overall principle. According to that principle, one alternative is unambiguously superior than another if and only if the second is never chosen when the first is available.

[14] See also Harrison (2019) for a detailed review in the context of choice of insurance products.

Just like the true preference approach, the quantitative intentional stance rejects BCI and keeps NI. That is to say, preferences are context-dependent and the adequate context to reveal the normative preferences of the individual is the lab (with the help of experimental tasks). It also keeps NCI, where the normative preferences estimated in the lab are taken to be context-independent (Harrison 2019). The quantitative intentional stance is still a third-person standpoint, where the ultimate judgement of what makes individuals better off belongs to the theoretician. The authors justify their approach pragmatically, where an economist is hired as a 'consultant' in order to advise her clients. As a result, NCS is rejected but with the explicit consent of the client who expresses her willingness to delegate her states of affairs to the one of the consultant.

Just like the choice-based criterion, the main problem with the quantitative intentional stance is its restricted range of applications (violation of Unrestricted domain), although it offers for those situations an operational measure of well-being. This approach remains however silent about more problematic cases with strong conflicts between the preferences of the selves. Such result echoes with the one of Steedman and Krause (1986), who show that preference integration into a single preference ordering requires the 'character' of the individual (i.e. the type of the function she uses to aggregate her different preferences) to be '*harmonic and sensible*' (p. 219 — their emphasis). That is, it implies a limited conflict between the preferences of the multiple selves.

4.3 Contractarian Approach: The First-Person Standpoint

Unlike the third-person standpoint, Sugden (2004, 2018a) argues that the theoretician is not entitled to make value judgements about individuals' preferences. According to the author, an individual should be seen as a continuing locus of responsibility, treating her past and future actions as her own, whether or not those actions were or will be what she would like them to be now (Sugden 2004, p. 1018). This quality of responsible person gives normative relevance to the judgement of the individual on her own actions. The correct normative standpoint is here not the third-person standpoint (the theoretician) but the *first-person* standpoint (the individual herself). Sugden's approach to normative economics can thus be identified as the 'view from somewhere'. The idea is that social arrangement (e.g. competitive markets) is based on the acceptability of each member of the society who is considered to be a potential party to an agreement or social contract. This *contractarian* approach shifts normative appraisal from individuals' well-being to *opportunity to choose*, irrespectively of what individuals' preferences are (i.e. well-ordered or not).

In this approach, it is up to individuals themselves to contract whether it is in their interests to opt for one alternative over another. Taking again the Asian disease problem, the contractarian approach consists in letting individuals choose as they prefer in their current situation — contrary to the third-person standpoint, which attributes a normative criterion external to the subject on what frame is the 'correct' one. Such 'view from somewhere' assumes that the individual is responsible for her own choices, and is thus in the best situation to judge what makes her better off. This normative approach then drops BCI and NCI and determine the current context as the adequate one for normative analysis. The consumer sovereignty principle in this approach can therefore be reformulated as follows.

Behavioural consumer sovereignty (BCS). $\forall C' \in \Gamma, NP_{C'} = BP_{C'}$

NCS imposes that the normative preferences of the individual are complete and integrated (because they correspond to the behavioural preferences, which are themselves complete and integrated). BCS however does not impose any constraint on the normative preferences of the individual. She must be able to satisfy any preference that she *might have*, knowing that those preferences may change depending on the context.

On the assumption of the continuing individual, we have $NP^* = BP^*$. From this perspective, there is no significant problem with the fact that individuals act on non-integrated preferences. What matters is not individuals' well-being but rather their available *opportunities*. This means that economists should not interfere with individuals' private preferences (how they choose, once their set of available alternatives is set) but rather ensure that individuals have access to the larger sets of opportunities.

It is however worthy to note that Sugden's (2018a) approach is a *defence* of the market (as explicitly acknowledged by the subtitle of the *Community of Advantage*). The author's main concern is that behavioural economics can cause problems to economists justifying market institutions based on welfare evaluations. As a response, he proposes an alternative approach to normative economics based on opportunity, which does not require the empirical validity of the neoclassical representation of preferences. This means that this approach remains however silent on relatively uncontroversial cases that could be highlighted by behavioural economists such as self-acknowledged failures of self-control (e.g. drug addiction). Sugden's (2017b) response is that such genuine problems of self-control are quite rare. That is, extreme situations such as heroin addiction are not comparable to more common impatient behaviours such as not respecting one's diet. A related issue is that the contractarian approach cannot disentangle between cases of adroit marketing (such as a baker who prominently displays her nicest cakes rather than offering them already wrapped in cellophane) and fraud or deception on behalf of firms. These are however unacceptable behaviours to the extent that they violate the rules of fair competition. For instance, the opportunity criterion cannot say anything on the legitimacy of using ambient scent in supermarkets as a strategy to induce different moods and desires (Akerlof and Shiller 2015).

4.4 A Contractualist Proposal: The Second-Person Standpoint

Rather than leaving the task of defining the correct context to the theoretician (as in the third-person standpoint) or to merely accept the current one (which can be set by firms, for the better or the worse), we propose that it should belong to the continuing individual to define her own normative preferences and to proceed to the integration of her preferences. This alternative standpoint emphasises the role of the dynamic processes involved in the construction of one's identity. We argue that this position can be justified by either appealing to Sugden's (2004, 2018a) opportunity criterion while considering the question of preference integration, or to the analogue of Sen's (2009) 'positional views' in his theory of justice.

4.4.1 The Second-Person Standpoint

When I is embedded in a context C she knows that her choices may also impact $I_{C'}$, $I_{C''}$, etc. Assuming those selves are part of the same continuing person I^*, we can argue — following Darwall (2006) — that each self I_C has a moral obligation regarding the other selves, which are part of the same 'community of selves'. The idea of the *second-person* standpoint is that the adequate question one should ask in order to assess an alternative is not the respect of an external principle (as implied by the third-person standpoint) or the preference of the individual alone in a specific state of affairs (as implied by the first-person standpoint). Instead, it is the perspectives of *others*, and more specifically the ones to which the individual is morally obliged.

Consider as an illustration a teacher who has not finished her lecture on time. She may either finish it by keeping her students extra minutes, or let the students leave the room. From the third-person standpoint, she should appeal to an external normative criterion to make her choice (e.g. that more knowledge is always preferable, or that she should always respect schedules). The first-person standpoint would leave her to make her own choices based on what she considers to be important at the moment. Endorsing the second-person standpoint however means that she should consider the problem from her own perspective (e.g. she has a preference to finish the lecture because she thinks it is fascinating) but also from the perspective of her students (e.g. a few of them may be genuinely interested, while many just want to leave as they stopped paying attention a long time ago), and also from the perspective of the other teacher waiting for the room. It is then by aggregating these different judgements that she is able to form her normative assessment.[15] In the context of multiple selves (and apart from cases of severe identity disorders) we may legitimately consider that each I_C is morally obliged towards the other selves $I_{C'}$.

The traditional approach to model this kind of problem would be to suggest that I^* has preferences over contexts, and would therefore identify one context as the 'preferred' one to form her normative preferences. For example, the teacher would have preferences over the adequate *viewpoint* on the problem: of herself, of the students, of the other teacher. This however implies an infinite regress. Indeed, unless we assume that I^* has some intellectual capacities to think about the problem in a totally detached manner and to form a context-independent preference over contexts (like the inner rational agent), such preference is necessarily expressed in a specific context. This means that we should probably abandon the framework in which preferences and individuals are conceptually tied and rather draw an explicit distinction between the levels of I_C and the reflexive individual I^*.

4.4.2 The 'View from *Manywhere*'

Recall some of the notations we introduced earlier when we defined the 'context'. We defined each element $x \in X$ as described by a list of properties, some of which are known to the individual, while others are not. So far, we have taken BP_C as the primitive of analysis for each self and wondered about the process of integration of those preferences

[15] According to Darwall (2006), Smith's (1759 [2010]) notion of empathy makes him 'one of the first philosophers of the "second person", if not the very first' (p. 46). Carrasco (2011, p. 549) similarly argues that Smith's impartial spectator is not third personal but second personal.

into normative preferences. Note that the definition of the context from the very beginning is linked to how *we*, theoreticians, represent the choice problem. This point is without consequence for the contractarian perspective since the normative preferences in a given context always correspond to the behavioural preferences in the same context. In other words, in this approach normative analysis could be conducted independently of the theoretician's definition of the 'context'. This point is however crucial for the welfarist approaches since the individual's well-being has to be assessed by the observer. As previously argued, this also poses significant challenges from a social choice perspective. If we now consider that the proper locus of normative appraisal is how the individual forms her own normative preferences, we should make abstraction for the moment of the notion of 'context', which is a construction in the observer's mind that is not necessarily attainable by the individual.

When evaluating different alternatives, the individual is aware of a certain set of properties that characterises those alternatives. Note that at this stage the individual can be fully aware of conflicting rankings of the possible alternatives. For example, in the Asian disease problem the individual can be aware of the gain and loss properties. But rather than being determined by the choice environment, the self (and how the individual perceives the choice problem) is chosen by the individual herself. It is therefore I^* who may choose whether she will privilege a property over another. We here follow Davis (2011), who argues that individuals are continually redefining their identity through reflection, which requires the condition of *individuation*. According to this condition, the individual has the capacity to critically reflect upon her evolving preferences. In line with Dold and Schubert (2018), our normative proposal states that rather than focusing on the satisfaction of I's behavioural preferences, economists should rather contribute to improving the process through which she forms her own normative preferences. The normative criterion we propose is therefore the following.

Awareness Criterion. *For a given choice situation X, I is better off if and only if her awareness set $\mathcal{P}_I$ increases*.

Increasing one's awareness set allows the individual to consider the choice problem under additional new perspectives. For example, I becomes aware of the loss frame in the Asian disease, while she was initially only aware of the gain frame. In our account, the correct perspective from which to look at a choice problem is the 'view from *many-where*'. Rather than being satisfied with her initial perspective of the problem ('view from somewhere') or than endorsing a supposedly omniscient perspective of the problem ('view from nowhere'), what matters is her ability to *accumulate and confront* many views from different perspectives. Note that we do not expect individuals to look at every possible perspective on a given problem (such 'view from everywhere' would be akin to the omniscient perspective of the 'view from nowhere') but rather that more perspectives and different opinions or judgements on the same problem contribute to improving the process of decision-making. In the case of the Asian disease problem, subjects tend to intuitively identify the riskless option as the right one in the gain frame and the risky option in the loss frame. But it is only by confronting these two views that individuals are able to form an 'enlightened' judgement — regardless of what they choose *in fine*.

At the practical level, our principle implies that the role of theoreticians — who are

supposed to have an awareness set $\mathcal{P}_E$ that is larger than the set of individuals — should be to educate individuals on the existence of other perspectives. For example, economists can teach individuals to reframe probabilities as natural frequencies, as suggested by proponents of the *boost* agenda (Grüne-Yanoff and Hertwig 2016). By increasing individuals' set of perspectives, the process of decision for the individual is improved under the standards of the awareness criterion. Just as in Sugden's (2018a) 'view from somewhere', we argue that the addressee of normative economics should be *individual citizens* and not the abstract 'social planner'.

We see two ways of justifying the awareness criterion. First, we can interpret this normative criterion as an application of Sugden's (2018a [Chap. 5]) 'individual opportunity criterion' to the process of preference formation. Indeed, if we endorse a dynamic view of personal identity we should consider that the individual's preferences do not necessarily *pre-exist* from the choice situation but are rather *progressively determined by the process of choice*. As Nozick (1981) puts it,

> 'The reasons [considered in deliberation] do not come with previously given precisely specified weights; the decision process is not one of discovering such precise weights but of assigning them.' (p. 294)

It is likely that such process is path-dependent and the individual's ability to make creative choices depends on her initial sets of representations of the world. If the individual is only aware of one way to look at the world (e.g. she always chooses the cheapest alternative) then her future opportunities to learn new preferences are very likely to be reduced (see Schubert's (2015) 'opportunity to learn' criterion). But if we value opportunity when considering choice sets and accept that one's identity is the result of an evolving process and critical reflection upon one's experiences, then opportunity also seems to be valuable when considering the sets of possible future identities (Buchanan 1979 [1999]; Dold 2018).

Another justification of our criterion is to draw a parallel with Sen's (2009) 'positional views' in his theory of justice. Baujard and Gilardone (2017, 2019) emphasise that Sen's theory of justice is 'poorly understood' and that it continues to raise some debates — such as the debate about the proper place of capabilities in his contribution. We suggest that one of the reasons of those discussions is the specific standpoint Sen proposes in his approach, which is second personal. Sen indeed rejects transcendental approaches relying on one pre-defined normative criterion (third personal standpoint). He also considers the question of adaptive preferences to be problematic, suggesting that first personal approaches may be invalid because one's current judgement about one's well-being may be influenced by one's deprivation. Baujard and Gilardone (2019) argue that the concept of 'positional views' is central in Sen's (2009) *Idea of Justice*. The concept is defined as 'an individual judgement towards any social state, considering objectively the context from which she or he is able to assess this social state' (Sen 2009, p. 3). What is important is that one's positional view may evolve if information from different positions is communicated.

For example, the individual's view on implementing a universal basic income is likely to depend on various elements that characterise her current position (positional parameters) such as being in a situation of poverty or not, being in a health condition that limits her opportunities of employment, etc. Positional views can be 'objectivised' because they

can be characterised by some meta-criteria that are open to public debate (e.g. being poor or not, being handicapped or not, etc.). Since such views '(1) may influence observation and (2) can apply to different persons' (Sen 1993, p. 127) they can constitute a relevant input for collective decision. We advance that it is only by confronting the many views of individuals from different positions that individuals can collectively form an enlightened judgement about a specific state of affairs. The relevant inputs for normative analysis are therefore the views of the various individuals that constitute the society.

Sugden (2006) suggests that Sen's (2009) position implies 'that "we", as ethical theorists, can claim to know better than some particular individual what is good for her' (p. 34). According to the author, it implies that theoreticians *in fine* impose as the social standard 'the kinds of lives that a majority of [our] fellow citizens, after reflective deliberation and open debate, judge to be valuable' (p. 40). However, Sen (2006) rejects this 'monstrous political philosophy' (p. 89). Sugden (2006) indeed favours a first personal approach to determine what preferences are valuable from the perspective of the individual. That is, from the perspective of the theoretician, any preference that the individual might have is valuable. In this matter, the author attributes to Sen (2006) a third personal approach. Sen (1970 [2017]) however argues that we have to look for a social mechanism such that in a matter of '"purely personal" choice [the individual's] preference should be precisely reflected by social preference' (p. 130). The emphasis is put on the *process* — a 'social mechanism' — through which individuals' views can be confronted, while guaranteeing *in fine* the respect of one's own preferences. It is true that theoreticians can have pre-defined ideas about what counts as the good life, but such views should only serve as inputs among others in the collective discussion. Sen's idea is indeed that *open public reasoning* constitutes the adequate social mechanism guaranteeing the formation of social preferences. It is about confronting the different positional views — first as a way to *widen* the informational basis of all the participants, and second as a way to create a 'greater sense of neighborhood' (Baujard and Gilardone 2019, p.13). In such second personal approach, we find a trace of Smithian sympathy and the idea of the impartial spectator (Bréban and Gilardone 2020).

Sen (2009) does not make any presumption about the outcome of public reasoning. Just as in our proposal, Sen's (2009) theory of justice shifts normative appraisal from outcomes to the process of choice. This means that there is no ready-made theory of what a 'good' society is (or what 'good' preferences are) but that there is a general approach (confronting the different views on the same question) that contributes to form collective judgements. If the theoretician turns out to be aware of some positional parameters (in our framework, of properties influencing the choice) then he ought to inform individuals about those properties. Whether individuals *in fine* take them into consideration is however not relevant to the theoretician.

4.5 Conclusion

In this chapter we argue that the RP directly echoes to social choice debates about preference aggregation, as the RP is concerned with how to integrate the conflicting preferences of one's multiple selves. We contrast the different approaches one could take to solve the RP. These can be resumed in the welfarist third-person standpoint (assessing individuals'

states of affairs from the standpoint of an external observer: the 'view from nowhere') and the contractarian first-person standpoint (leaving individuals be the own judge of their well-being: the 'view from somewhere'). Since Arrow's (1951 [2012]) impossibility theorem can be adapted to this framework, we highlight fundamental difficulties when it comes to defining an integrated welfare function. Because of these theoretical issues, we propose an alternative standpoint for normative economics in which normative assessments are made possible through the individual's evolving identity. We label this approach the 'view from *manywhere*'.

These three standpoints of the RP (and more generally of how to make normative assessments) also provide different policy guidance. By conferring a significant role to the external observer, the third-person standpoint defines what may count as 'welfare-relevant' by characterising the 'correct' context. This normative approach can justify direct policy interventions such as *nudges* (Thaler and Sunstein 2009) or more traditional forms of interventions such as taxes. The first-person standpoint typically takes individuals' preferences as 'protected', and rejects any policy intervention aiming at interfering with the expression of individuals' freedom of choice. In this normative approach individuals themselves are responsible for their own choices. This implies that public policies should be limited to guarantee the respect of the rules of fair competition. On the other hand, the second-person standpoint focuses on the process of preference formation and decision-making. It offers a rationale for more 'educational' policies. The aim here is to foster individuals' abilities to critically reflect on their own preferences and increase their opportunities to learn new preferences. By contrast with *nudges*, *boosts* — defined as policy interventions that help increase individuals' informational basis (Grüne-Yanoff and Hertwig 2016) — are well aligned with the normative standpoint we propose.

CHAPTER 5

Identity, Personal Persistence and Normative Economics

Abstract

Multiple selves is a conventional assumption in behavioural welfare economics for modelling intrapersonal well-being. Yet an important question is which self has normative authority over the other. In this chapter, we tackle this ethical question through the ontological question of personal persistence: *what does it take for an individual to persist from one time to another?* We review the main theories of personal persistence offered in analytic philosophy and discuss the philosophical problems related to the alternative unified assumptions of the self offered in the critical literature of behavioural welfare economics. We discuss two main issues. First, most of the authors defending a unified account of the self in normative economics tend to consider the question of identity over time from an *ethical* viewpoint ('which theory of identity best fits with our idealistic picture of what a person should be?') but not from an *ontological* viewpoint ('what makes an individual numerically identical from one time to another?'). We argue that the ethical viewpoint is misleading because it reduces the question of personal persistence ('what makes an individual identical from one time to another?') to the question of *personhood* ('what does it take for something to be a person?'). Second, we discuss the fact that the alternative assumptions of the unified self endorsed in the critical literature of behavioural welfare economics assume the *narrative* view of personal persistence. Because of its many philosophical objections, we however argue that the narrative view cannot provide a satisfying account of identity. We conclude that in order to solve some important ethical issues of identity in normative economics, one has deep interest in looking for ontological approaches that better define what makes an individual persist through time.

Acknowledgements. Joint work with Valerio Buonomo (*Università degli Studi di Milano*) and submitted version to *Economics and Philosophy* (July 2020). We thank John Davis, Cyril Hédoin and Marya Schechtman for useful feedbacks on early drafts of this chapter. All mistakes remain ours.

5.0 Introduction

With the growing interest of behavioural economics towards the evaluation, recommendation and prescription of public policy (Thaler and Sunstein 2003, 2009), a conventional assumption is to consider each individual as being composed of at least two selves: a far-sighted 'planner' and a myopic 'doer' (Thaler and Shefrin 1981), 'hot' and 'cold' states (Camerer et al. 2003), or an automatic system 1 and a reflective system 2 (Kahneman 2011). In a nutshell, the myopic doer/hot states/system 1 are the ones in which decisions are driven by fast thinking or made in the heat of the moment (e.g. eating a cake), whereas the far-sighted planner/cold states/system 2 are the ones in which decisions are driven by slow thinking or made reasonably (e.g. avoiding the temptation of eating a cake). This dual conception assumes that individuals make decisions they may later regret, and that their normative authority is better located in the far-sighted planner, cold states or system 2.[1] However, such assumption of multiple selves comes up with a major difficulty from a philosophical viewpoint: how to locate individual normative authority when it is left unclear which of the preferences of the many possible selves are truly normatively relevant? In a recent appraisal about the value of individual autonomy in libertarian paternalism, Sunstein (2019) acknowledges this concerning problem when he raises doubts regarding the arbitrariness of making ethical judgements about which self has normative authority over the other. In his words,

> 'What doers do might be one of the most significant and best experiences of their lives — even if they would have chosen otherwise in advance and perhaps even if they regret it afterwards. ... Why does John or Edith deserve authority at Time 1 or Time 3, rather than Time 2? What makes either of their views authoritative or authentic, rather than the choice at Time 2?' (pp. 69-75)

Similarly, Kahneman (1994) raises important ethical concern when observing conflicting evaluations of patients during and after being subject to painful experiments:

> 'The history of an individual through time can be described as a succession of separate selves, which may have incompatible preferences, and may take decisions that affect subsequent selves ... Which of these selves should be granted authority over outcomes that will be experienced in the future?' (p. 33)

This philosophical problem is particularly salient when one makes the assumption of multiple selves. Indeed, ethical questions such as what the relationship between different selves are, or whether selves differ in the same ways individuals differ seem to be merely unavoidable from a philosophical perspective. Yet the multiple selves assumption is not the only way to describe intertemporal choice. For various reasons, some authors are sceptical about the multiple selves assumption and instead propose an alternative assumption of the unified self in which the individual is represented as a *unified* being through time.

[1] Those types of decision are mostly represented by the known psychological phenomenon of self-control failure. Self-control failure is explained by several models of decision-making, such as quasi-hyperbolic time discounting — which encapsulates the idea that individuals have present-bias preferences (Laibson 1997) — or an axiomatic foundation of the ranking of commitment, of temptation, and cost of self-control (Gul and Pesendorfer 2001). In the present chapter, we are however not exclusively concerned with self-control failure but with any kind of intertemporal choice that may affect one's well-being, e.g. how much to save for retirement, how to invest, whether to buy a house, whether to have children or whom to marry (Loewenstein and Thaler 1989).

Among these authors, Sugden (2004) drops the multiple selves assumption in his early reconstruction of welfare economics without the concept of preference by instead considering the individual as 'a continuing locus of responsibility ... to the extent that, at each moment in her life, she identifies with her own actions, past, present, and future' (p. 1018). In his theory of identity in economics, Davis (2011) criticises the fragmentation and dissolution of the individual into multiple selves and argues that a criterion of identity should satisfy what he calls the *individuation* and *reidentification* criteria (to be discussed below). The author argues that a unified concept of the self based on a capability approach (Nussbaum and Sen 1993) best satisfies his two criteria. In a similar trend line, Hédoin (2015) holds a narrative view of identity based on Korsgaard's (1989) definition of agency, which he claims to palliate the problem that selves cannot be attributed moral responsibility. Also, the philosophical issues related to identity in behavioural welfare economics lead Dold and Schubert (2018) to suggest that 'the dualistic concepts of the individual should be abandoned in favor of a notion of a unified self that is constituted by its capacity to learn and reflect upon new preferences on a continuous basis' (p. 221). Although appealing, the main problem with most of these alternative views of the unified self is that they are based on ethical assumptions rather than on ontological theories about *what makes an individual truly one*.[2] But arguing how an individual remains the same from one period to another inevitably results in debating about the ontological status of what it takes for an individual to persist over time — a terminology used in analytic philosophy for what could characterise 'economic agency' in economics.[3]

The aim of this chapter is to clarify the philosophical difficulties of endorsing the assumption of the unified self as an alternative to the multiple selves assumption in normative economics. To our knowledge, those difficulties have not yet been investigated. The literature of identity-and-economics is of course extensive and goes beyond the philosophical problems of behavioural welfare economics (Akerlof and Kranton 2000, 2010; Davis 2003; Kirman and Teschl 2004). We however do not exclude the possibility of our study to be relevant to the broader program of identity-and-economics, nor to normative economics in general. By behavioural welfare economics (BWE), we refer to the literature which boils down to reinterpreting normative economics when individuals have incoherent preferences.[4] We include in this literature Camerer et al. (2003); Thaler and Sunstein (2003, 2009); Bernheim and Rangel (2007, 2009) and Bernheim (2009, 2016).[5] The ethical problems of BWE related to identity have particularly been the matter of interest of Ferey (2011), Lecouteux (2015a) and Hédoin (2015) in relation to Parfit's (1984) theory of identity. Although we have no particular objection with the philosophical

[2] An exception is Davis (2011), to be discussed below.

[3] The notion of agency is actually complex in economics and may refer to different meanings (see Sen (1987 [2003]) for a definition in terms of personhood). As Hédoin (2020) puts it, we may represent agency as a combination of two sets of properties: *rationality* properties (to be characterised as rational in some sense) and *identity* properties (to be the same individual in space and time). We here only refer to the latter, not the former.

[4] We here avoid to focus on the reasons why preferences can be incoherent (e.g. present bias, framing or loss aversion) as they would add nothing relevant to the goal of the present chapter. That is, we merely consider incoherent preferences to be synonymous with *preference reversals*: the case in which an individual prefers A to B and B to A at two different times.

[5] There is also the adversarial anti-welfarist approach of Sugden (2004, 2018a) that we do not include in behavioural welfare economics because it is merely not welfarist. For a general review of the normative program which aims at 'reconciling normative and behavioural economics' in response to the observation that individuals have incoherent preferences, see McQuillin and Sugden (2012).

issues of the multiple selves assumption discussed by these authors, we argue that holding the unified self assumption is nonetheless no less problematic as it requires economists to justify the underlying personal persistence theory on which their conception of identity is based on. We introduce the literature of personal persistence into this 'multiple selves *versus* unified self' debate in normative economics which, we argue, is an important matter of interest for two reasons.

First, personal persistence provides what we judge to be a better framework for discussing the relationship between identity and ethics. The point is that many of the alternative assumptions of the unified self proposed in the literature do not take an *ontological* but an *ethical* viewpoint of the criterion of identity. That is, they start from an idealistic picture of the individual based on a concept of personhood and then make it an assumption to palliate the philosophical issues of multiple selves. We however argue that this viewpoint is methodologically misleading since the question of identity over time cannot be reduced to the question of personhood ('what does it take for something to be a person?'). Importantly, focusing on the ontological question of personal persistence does not set aside the ethical concerns one can make within BWE. On the contrary, ethical concerns about identity — such as the ones discussed by Sunstein (2019) and Kahneman (1994) — can be informed by the ontological account of identity. In other terms, we suggest an 'ontology first' approach, where issues coming from the ontological debate on personal identity can support, and eventually steer some discussions in the identity debate in normative economics.

Second, the literature of personal persistence — yet unknown to the literature of identity-and-economics — enlightens the difficulties associated with the current alternative assumptions of the unified self proposed in the critical literature of BWE. Particularly, some economists who are sceptical about the multiple selves assumption endorse the *narrative* view of the unified self in order to avoid the philosophical problems associated with multiple selves (Sugden 2004; Davis 2011; Hédoin 2015; Dold and Schubert 2018). We however bring about some philosophical arguments that give us good reasons to be also sceptical about the narrative view of identity.

Because of these two issues we discuss throughout the chapter, our main argument is that to solve ethical problems of identity in normative economics does not necessarily require to make ethical judgements about the constitution of personhood. Instead, one has good reason to proceed in a different way, focusing first of all on the ontological criterion of identity over time — an enquiry we tackle head-on in the present chapter. The rest of the chapter is organised as follows. Section 5.1 introduces an example of the identity problem in ethics applied to BWE and contrasts our ontological approach of identity to related literature. Section 5.2 presents the framework of personal persistence by formalising the criterion of identity over time. Section 5.3 reviews the main theories of personal persistence offered in analytic philosophy. We show that most of the alternative unified self assumptions proposed in the critical literature of BWE cope with the narrative view of personal persistence. We argue that those assumptions of the unified self are no less criticisable than the multiple selves assumption because (i) they tend to define identity from an ethical viewpoint and (ii) because the narrative view is philosophically problematic. Section 5.4 concludes.

Two useful clarifications should be stated early. First, we deliberately (and whenever needed) privilege the term *personal persistence* instead of identity in order to use a more accurate terminology — albeit it is not absurd to consider 'identity over time' as a synonym of 'persistence'. One reason is that identity is a potential result rather than an assumption in the literature of personal persistence. After a careful investigation of how temporal selves of individuals are related to each other, identity may, after all, not be what matters for personal persistence.[6] Second, *temporal* selves instead of multiple selves is also a privileged terminology because we are here only interested in selves which differ with respect to their temporality. That being said, we do not deny that the concepts of doers/hot states/system 1, on the one hand, and planners/cold states/system 2, on the other hand, may be understood as coexisting at any point of time. For example, the doer tells the individual to eat the tasty cake, while the planner prevents her at the same time from doing it. But in order to be consistent with models of intertemporal choice, we here consider individual behaviour to be a matter of preference reversal over time, where the preference of one self overrules the preference of the other self at the point where the decision is being made.

5.1 The Ethical Problem of Identity

Assuming the view of multiple selves, an individual I can be considered over time as a collection of a finite number of temporal selves $\{s_1, ..., s_n\}$, where s_i is a temporal self which exists at a given time t_i. Let $i = \{1, ..., 100\}$ be the index of time. Imagine that at t_1, s_1 preferred A to B, whereas now, at t_{100}, s_{100} changes her mind and prefers B to A. As underlined by Sunstein (2019) and Kahneman (1994), one central matter of concern in intertemporal choice is which of the two (or of the many other) selves has/have normative authority.[7] We first consider several intuitive ethical rules in turn and argue that they all suffer from being arbitrary. We then argue that the ethical problem of identity constitutes a practical burden for economists, particularly when these rules are based on empirical evidence. This brings us to formulate the ethical problem of identity in the ontological framework of personal persistence, which we introduce in Section 5.2.

5.1.1 The Present Rule

A first rule would state that s_{100} overrules s_1 because what matters is what happens *now*. That is, an individual is the master of her own well-being at each of her present temporal part. This rule gives the present self full responsibility about her own actions. In this sense, it is well aligned with the liberal tradition of the consumer sovereignty principle in economics, where each s_i is the best judge of her own well-being at t_i. Instead of looking for which preferences count as normatively relevant, this rule is compatible with Sugden's

[6] This is for example the conclusion of Parfit (1984, p. 215), according to which identity does not matter in persistence in terms of survival, as identity and persistence/survival involve different kind of relations. While identity is a one-to-one relation, persistence/survival is a one-to-many relation (in terms, for instance, of mental continuity). These points are however of marginal importance for our present study because we exclusively focus on the *unified* account of the self. For an introduction to Parfit's account of personal survival, see Shoemaker (2019, sec. 2.5). For critical appraisals of the implication of Parfit's theory of personal persistence to BWE, see Ferey (2011, pp. 746-747), Lecouteux (2015a, pp. 403-407) and Hédoin (2015, pp. 98-102).

[7] By 'normative authority' we mean 'moral responsibility on all the other selves'. This example is a version of McMahan (2002, p. 497) formulated on our own.

(2004, 2018a) view that economists should better promote institutional arrangements so that individuals can seek what they want — disregarding how incoherent their preferences are.

5.1.2 The Priority Rule

A second rule would state that what matters is the preference of the *first* temporal self. That is, if s_1 expresses a preference for A over B, then s_1 overrules s_{100}. This rule is likely to hold on the important condition that s_1 contracts with her (not yet existing) future selves. The contract would specify that s_1 takes full responsibility of the consequences of her preference for s_{100}, no matter what they are. Such rule would however be endorsed by BWE only if it appears that s_1 is the far-sighted planner, cold state or system 2. But if not, s_1 is making a mistake and her preference would not be considered as normatively relevant.

The main issue with both of these rules is that they seem quite arbitrary from an external viewpoint, particularly when we have no objective criterion to determine what makes the normative authority of s_1 more important than s_{100} (and conversely). As Sunstein (2019) puts it, 'there is no alternative to resorting to some kind of external standard, involving a judgment about what makes the chooser's life better, all things considered' (p. 79).

5.1.3 The Objectivist Rule

A third alternative would then aim at shouldering this external standard by trying to determine with 'reasonable' assumptions an objective criterion that would state which self (or selves) has/have normative authority — no matter their past, present and future status. There may be at least two ways to define such objective criterion.

The Majority Criterion

One may say that if most of the selves among the finite number of all selves prefer A to B — say, $\{s_1, ..., s_{51}\}$ but not $\{s_{52}, ..., s_{100}\}$ — then fifty-one selves against forty-nine overrules whatever s_{100} states, and then the preference of A over B should be taken as the one which is normatively relevant. This view is explicitly endorsed by Thaler and Sunstein (2003, p. 178), who argue that the social planner should choose a choice architecture based on the majority of individuals' expressed preferences. But if it appears that the fifty-one selves are the myopic doers and the forty-nine other selves are the far-sighted planners, economists may not consider this criterion to be reasonable.

Specifically, the majority selves should be expected to have a minimal form of rationality so that economists may reasonably think that their judgement are 'enlightened'. Thaler and Sunstein (2003) self-recognise the limits of the majority criterion, stating that the majority's choice may simply not be sufficiently informed, and that those aggregated choices may not promote the majority's well-being. Another important issue with the majority criterion is that it assumes that selves are equally weighted, but it does not have to be so.

The Weighting Criterion

One may say that one temporal self has more normative authority over the others, and the issue is to know which of the finite number of all selves is the 'supreme' one. For example, assume only s_{30} and s_{100} prefer B to A (among all the remaining selves), and we discover (by some knowledge) that s_{30} has supreme normative authority. Based on this objectivist rule, we should therefore account for the preference of s_{30}. This view is endorsed by Bernheim and Rangel (2007, 2009). Albeit the authors do not explicitly refer to a weighting criterion, their extension of the revealed preference framework to welfare analysis requires minimal guidance to individual well-being through choice data. In their account, only the 'fully rational' self counts for welfare analysis, even if most of the remaining selves prefer A to B. Again, this view is however not unproblematic.

We may obviously question what the good reasons are to make us believe that one self should be given more weight at one period of her life instead of another. It seems that cognitive capacities are here important: some temporal selves could be eliminated from the possibility of having any normative authority — typically the ones who belong to childhood. But accordingly, we would also need to determine what is the 'mature and reasonable' period of one's life. Assume now that empirical evidence could tell us which period of one's life tends to be associated with one's true preference, e.g. all temporal selves included in the set $[s_{30}, ..., s_{40}]$, and assume that we could somehow determine such interval. It would nonetheless also require that the preferences of the selves which belong to this interval remain relatively stable. But recall that the initial problem of BWE is to find a way to reinterpret normative economics when individuals have incoherent preferences. This means that if one self which belongs to the set $[s_{30}, ..., s_{40}]$ states incoherent preferences, we are left at the same point.

There can be, again, objective criteria to overcome that issue, e.g. consider the mean value of the interval (here t_{35}) as the instant with the temporal self having normative authority. Yet the issue with any kind of objectivist criterion is that it rules out the possibility of idiosyncratic preferences. In other terms, a general rule of individual well-being waives the possibility that individuals can perceive their 'mature and reasonable' period of life differently. For example, some may make more sense to their life as a whole at s_{67}, while others at s_{24}.

5.1.4 The Ontological Viewpoint on Personal Persistence

The ethical problem of identity in BWE seems, from a practical point of view, quite a burden for economists to solve — particularly when they have no expertise in determining on which general criterion they can locate normative authority. Perhaps only economists' ethical judgements can help them out, but those ethical judgements are far from being self-evident and subject to consensual agreement. For example, Bernheim (2016) fairly underlines the problem of 'heavily value-laden language' (pp. 38-39) such as 'present bias' and 'self-control problems', which assume that individuals have unitary preferences and equate well-being with exponential-discounted utility. Bernheim (2009) also emphasises that true happiness might be interpreted as living at the moment. That is to say, there is *a priori* no reason to consider the present rule as less important than any other rule. In addition, it is not impossible that ethical judgements made by economists, which are expressed in terms of preferences, are also subject to incoherence from one time to another.

A social choice alternative that we do not engage into in the present chapter is to offer an ethical account of how to aggregate well-being over the temporal selves.[8] In his discussion related to the philosophical issues of identity in BWE, Hédoin (2015, pp. 84-88) specifically tackles the ethical problem of identity from this perspective by formulating a social welfare function of BWE. According to this function, the social planner maximises the weighted sum of the selves' utilities of a given population with respect to an exogenous weighting parameter. The author points out the difficulty of knowing the weight of each self in the decision, especially when there are no other alternatives than making ethical judgements about which selves' decision is considered to be more normatively relevant than another (p. 89). This social choice alternative is also well recognised by Sunstein (2019), who notes that 'the experiencing self might have too little regard for the remembering self, but the converse is also true. It is not clear that either deserves priority. To know, we might have to make some moral judgments, or offer some account of how to aggregate well-being over time' (p. 76). Aggregating well-being over time would however ultimately yield to the arbitrary ethical rules previously discussed.[9] This is particularly the difficulty we aim to avoid by not focusing on the ethical problem of identity from a *social choice* perspective (how to aggregate individual well-being at the intrapersonal level) but from a *personal persistence* perspective (what makes an individual one over time). Although we are very much sympathetic towards the social choice alternative, we here take the ethical issue of identity at its roots by focusing on the *criterion of identity over time*.

The way we see the ethical identity problem introduced above, any answer that comes up to determining the moral authority of an individual requires to make essential reference to personal persistence. To put it differently, we defend what we believe to be the receivable view that what makes an individual morally responsible for her own action at a given time is a question that cannot be answered without first asking ourselves what makes that same individual persist over time.[10] That is to say, *the individual I can be held responsible for her past and future actions only if I is the same individual from one time to another*. Importantly, contrary to the critical literature of BWE which grounds identity on ethical claims (to be discussed below), we shall add that we do not say anything on what is required to have moral responsibility. Otherwise our approach would take the same path as this related literature. We specifically aim to avoid any ethical stance that assumes or defines the concept of morally responsible individual in order not to bias our enquiry of what makes an individual persist over time. This 'unbiased stance' is required if we do not want to first make (arbitrary) ethical claims about what morally responsible individuals are, and then conclude on what makes the individual persist from one time to another. That is, we aim to take the reverse stance of Sugden (2004), Hédoin (2015) and Dold and Schubert (2018) in their account of identity. We first ask what makes an individual persisting over time *and then* this ontological enquiry will help us comment on the ethical problems faced in normative economics. Accordingly, we will from now on use the term 'person' to define an individual who has moral responsibility, and we shall insist that contrary to this related literature, we do not say anything on what is required

[8] We deal with this task in Chapter 4 by making the assumption of multiple selves.

[9] We of course do not mean those rules to be exhaustive. Any other rule, for it has reasons to be justified, is receivable.

[10] We say 'receivable' because some philosophers object the view that any ontological question about personal persistence is relevant to our practical ethical concerns (Rovane 1998; Conee 1999).

to have moral responsibility.[11]

In addition, it would be in our view misleading to look for empirical evidence if one wants to define personal identity. Empirical evidence would provide what is required to rationally claim that a person persists through time, e.g. observing the physical properties of an individual through time. But empirical evidence is in fact neither necessary for personal identity over time — one might change some physical properties and/or forget some of her past events, but nonetheless persist across time — nor sufficient for them — we might imagine the case in which there is another person, who resembles a friend of ours, and is aware of the most important events of her life. In other terms, empirical evidence provides answers to the so-called evidence question 'how do we find out whether a person at one time is numerically identical to a person at another time?' (i.e. 'what evidence do we need to maintain that a person we see today is or is not the same person we saw yesterday?'), but not to the ontological question 'what are the constitutive conditions of personal identity over time?' (i.e. 'what does it take for a person to persist from one time to another?').

In order to avoid (i) intuitive reasoning that relies on common sense about which self has normative authority over the other(s), (ii) a social choice approach which would consist in proposing ethical rules to aggregate intrapersonal well-being, and (iii) the empirical limits of defining personal identity, we propose to formulate the identity problem of BWE within the framework of personal persistence from an ontological viewpoint. Contrary to the evidence viewpoint, we ask ourselves what are the constitutive conditions of personal identity over time instead of what are the factual evidence to conclude that a person persists over time. In the next section, we introduce such framework (Buonomo 2018) and then discuss in Section 5.3 four main theories that aim at explaining the unification of an individual's temporal selves.

5.2 The Criterion of Identity over Time

The problem of personal persistence consists in focusing on the *criteria of identity over time*. A criterion of personal identity over time can be defined as the completion Φ of the following schema.

> Let x be an entity that exists at time t_i and y an entity that exists at time t_j
> $\forall i \neq j$ and $Px \vee Py$ then $x = y$ if and only if $\Phi(x, y)$,

where $=$ is the relation of numerical identity, P is the property of being a person, and Φ is the constitutive condition whereby the identity of x and y is determined. Numerical identity is to be distinguished with qualitative identity. Two things are 'qualitatively identical' if they share the same properties (e.g. two identical chapters), whereas they are 'numerically identical' if they are one thing, and not more than one (e.g. the chapter

[11] In any case, one may simply argue that to take a position on what it takes for a living entity to have moral properties is a question that belongs to the land of metaethics. We however elaborate in the next section our position regarding the relevant question of whether personal persistence, when related to ethics, does not ultimately reduce to the question of personhood ('what does it take for something to be a person?').

you are reading right now). More generally, we can say that two things are qualitatively identical if and only if they exactly resemble each other, whereas they are numerically identical if and only if they are one and the same thing. By the logical disjunction $\vee$, we mean that we do not impose the condition that both x and y remain a person from t_i to t_j. This is an important distinction to be emphasised, as the critical literature of BWE typically assumes that (i) a person exists through time and (ii) her relationship with her future selves is necessarily related to another person. Such conception of identity is shared by Sugden (2004), who defines a responsible agent as a human being who,

> 'treats her past actions as her own, whether or not they were what she now desires them to have been. Similarly, she treats her future actions as her own, even if she does not yet know what they will be, and whether or not she expects them to be what she now desires them to be.' (p. 1018)

Similarly, Hédoin (2015) defines a responsible agent as a human being which is,

> 'responsible for all her actions and is interested in the consequences not only of her present action but also in the consequences of the future ones.' (p. 99)

These conceptions of identity already assume that there is no sense to argue about a person being an embryo in the past or a human vegetable in the future since they see identity as a relationship between persons defined as e.g. rational thinkers (as in the psychological view presented below) or as a psychological unity defined by a narrative (as in the narrative view presented below). In our present framework, we however do not want to make such essentialist assumption about persons because it would tend to reduce the question of personal persistence to either (i) the question of the ontological nature of persons ('what are we really?'), or to (ii) the question of the concept of personhood ('what does it take for something to be a person?') or even to (iii) the question about the biographical identity of persons ('who am I?').[12]

We recognise that the relationship of personal persistence with these three questions may be intimately linked when ethics is involved. Some may argue that for the sake of our practical ethical concerns, individuals persist over time by some psychological relation between their moral properties, which eventually constitute their identity. They may think that ethics inevitably forces us to endorse an essentialist personhood account of identity, as it seems irrelevant to be concerned with embryos or human vegetables — who by nature do not have the ability to produce any thought.

But to argue that identity presupposes morality (or any kind of psychological relation) seems a very strong claim. In fact, early conditions of identity related to moral properties would make us think that a concept of unified self necessarily has to be either psychological or narrative. We particularly think of the following identity conditions proposed by Hédoin (2015) based on Korsgaard's (1989) representation of agency.

- *Boundary condition.* I can be relatively easily identified as being I through her agency, including intertemporal agency.
- *Narrative condition.* I thinks of herself as a unit of agency and can make sense of the continuity of her decisions made in the past and the decisions she is thinking to make in the future.

[12] Some philosophers of personal persistence do impose the condition that $Px \wedge Py$ instead of $Px \vee Py$ (Swinburne 1984; Lowe 2012), which can be referred to as 'personal essentialism'.

The narrative condition here presented cannot be an assumption of personal persistence since nothing *a priori* tells us what constitutes the relationship between different temporal selves. This leads us to impose the following informative condition of a criterion of identity. A condition of identity Φ is informative if,

- **Non-triviality.** It has a different meaning from, or at least is not logically equivalent to the identity it constitutes.
- **Non-redundancy.** It should be logically possible that x and y do not satisfy Φ.
- **Non-identity-involving.** It does not presuppose the identity it should demonstrate.

Otherwise it is uninformative. For example, the statement that '$x = y$ if and only if they are the same entity' is trivial and identity-involving because it has the same meaning and presupposes the identity it ought to demonstrate. In contrast, the narrative view, which states that '$x = y$ if and only if they can make sense of their psychological continuity' is not trivial, nor redundant, nor identity-involving.

In the next section we review and discuss the main theories of personal persistence offered in the literature of analytic philosophy in the light of the framework we just introduced. The presentation of those theories nurtures the 'multiple selves *versus* unified self' debate in normative economics in the way that it exposes several possible views of the unified self, which are not necessarily narrative.[13] Importantly, knowing that the narrative view is dominant in the critical literature of BWE, we discuss some of its philosophical problems. Because of the objections associated with most of the theories of personal persistence we present, we end up suggesting that economists interested in the ethical identity problem of BWE should better allocate their efforts in making correct assumption about the ontology of individuals rather than trying to make correct assumption about personhood.

5.3 Theories of Personal Persistence

5.3.1 The Psychological View

The psychological view claims that an individual is identical over time in virtue of some psychological aspects such as memories, intentions, beliefs, goals, desires, and similarity of character (Parfit 1984, p. 207). That is,

> Let x be an entity that exists at time t_i and y an entity that exists at time t_j $\forall i \neq j$ and $Px \vee Py$ then $x = y$ if and only if x and y are connected by some given psychological relations.

This view has had by far the most advocates, mainly because of its practical appeal: how can y be responsible for the actions of x if she is not the inheritor of x's psychology?

[13] The next section is largely based on the taxonomy of Shoemaker (2019), who reviews the main theories of personal persistence and discusses their various normative implications. Whenever needed, we associate each view of the unified self proposed in the critical literature of BWE with the underlying theory of personal persistence it endorses.

Yet a generalised issue concerning the psychological criterion is that it seems to imply that personhood is one's essence — i.e. that an individual could not exist without being a person. But as previously argued, the question of personal persistence cannot be reduced to the question of personhood.[14]

Another concerning issue is that the notion of 'psychological' is not well specified as it may contain many aspects such as memories, intentions, beliefs, goals, desires — and importantly to economics — preferences. But the main concern of BWE is specifically about finding a normative approach consistent with behavioural economics when some psychological aspects of individuals, principally individuals' preferences, are incoherent for reasons economists do not fully understand (Bernheim 2016, p. 13). A continuity of incoherent preferences would imply to justify how those incoherent preferences are actually continuous, which seems a challenge one cannot face without relying on some essentialist assumptions. Those essentialist assumptions could be e.g. the existence of a far-sighted planner, or the existence of true preference: an essential property of the 'inner rational agent' (Infante, Lecouteux, and Sugden 2016a). But can we assume that one's identity is located in one's psychological property that neither behavioural economists nor neuroeconomists are able to locate?

Furthermore, it seems presumptuous to argue that one is constituted by an inner rational agent, which is the source of one's normative authority (and potentially also her identity), but that this individual cannot make way for other 'psychological roles' when she makes a decision, e.g. making a decision as a parent or a wife.[15] It is thus not surprising that no one has so far proposed a convincing account of the psychology of the inner rational agent. This rational agent supposedly has true (or latent) preferences that are

[14]One may object that this is not necessarily true for all psychological accounts of persistence, as it is the case for the Parfitian reductionist account. Indeed, one may argue that Parfit (1984) endorses a reductionist view of personhood, which should not be confused with the assumption that personal identity is a primitive. However, using this Parfitian counterexample as a general defence for psychological accounts of personal identity is (to us) misleading. This is because the Parfitian account is a very specific and non-standard account of personal persistence, which is characterised by the (very non-standard) rejection of the assimilation between 'personal identity' and 'personal persistence' — commonly summarised in Parfit's famous sentence 'identity does not matter for survival'. Given this account, the ethical claim of personal identity is prioritised, whereas the ontological question follows. Our intention here is not to discuss Parfit's approach, but rather to reject the use of Parfit's reductionist psychological approach as the standard approach for psychological views on personal identity. We stress that Parfit provides a very specific and not generalisable argument to face the first objection against psychological views. On the revisionary aspects of Parfit's theory, see Rovane (1998, p. 11) and Martin (1998, p. 15).

[15]We suspect some readers to answer that point by saying that a decision of a parent or a wife merely goes out of the scope of economic theory, and that it is therefore pointless to talk about a behaviour which is not even taken care of by the theory. If seriously claimed, we believe this point to be misleading, considering that leading behavioural economists take any sort of behaviour to be explained by intertemporal choice, such as how much schooling to obtain, whom to marry, or whether to have children (Loewenstein and Thaler 1989, p. 181). Thaler and Sunstein (2009) consider any kind of life situation examples to justify libertarian paternalism, such as avoiding the temptation of eating too much of the cashew bowl nuts before dinner (p. 40). Camerer et al. (2003, pp. 1244-1245) even consider the decision of committing suicide as a case for policymaking in their proposition of asymmetric paternalism. All these examples concern intertemporal choices that go beyond the archaic delimitation of economics to a limited set of decisions such as consumption, production, saving and investment. Although we do not particularly support the rhetoric of libertarian paternalism on justifying *nudging* from any kind of life decision such as eating from the cashew bowl nuts, we seriously assume the view that economic theory can explain any kind of choice which involves intertemporality.

accessible under conditions where she is undistorted from cognitive biases (Sugden 2015; Lecouteux 2016). But it remains a mystery whether those true preferences are actually produced or assumed to exist exogenously — as in the neoclassical consumer choice theory.

Arguably, such psychological view may miss something that the critical authors of BWE have argued to be important for identity: the *meaning* one attributes to the relation of her psychological relations (e.g. desires, intentions, life goals). In his reconstruction of normative economics without the concept of preference, Sugden (2018a) argues that it is only required to assume that an individual is a 'responsible agent', who can give a continuous meaning to each of her own actions at any given period of her life. This seems to avoid the practical burden of justifying the circumstances under which the selves have normative authority — that is, the circumstances under which they do not make cognitive mistakes.[16]

Sugden's (2004) view is also similar to the way Hédoin (2015) and Dold and Schubert (2018) interpret identity in normative economics. We discuss their narrative view of identity below. Before doing so, let us briefly introduce another approach that is compatible with our position that personal persistence cannot be reduced to personhood, but which, at the same time, claims that identity is not a matter of a psychological relation (an assumption we are so far quite sympathetic to).

5.3.2 The Physical View

Philosophers who are unsatisfied with the psychological view argue that it should not be a matter of fact that personhood is the essence of an individual, simply because it is hard to deny that an embryo who becomes an individual and then a human vegetable is not the same individual (Olson 1997; Hershenov 2005). Instead, it would perhaps be more convincing to define a continuous individual with respect to her physical properties. This view fits well with our current understanding of the human thought being reduced to physical features, such as brain cells and the complex connexions of the neural system. According to the physical view,

> Let x be an entity that exists at time t_i and y an entity that exists at time t_j
> $\forall i \neq j$ and $Px \vee Py$ then $x = y$ if and only if
> x and y are connected by some given physical relations.

Physical relations are not necessarily located in the brain. More generally, the physical view states that physical continuity, which constitutes the biological organism of a human animal, is the constitutive condition for personal identity over time, and then her persistence.

The physical view seems nonetheless far less appealing from an ethical viewpoint because it seems irrelevant to locate identity in a physical property that has *per se* no function of minimal reasoning or consciousness. But advocates of the physical view typically argue for a biological continuity between all stages of the body as a whole, e.g. from an embryo to a rotten skeleton. For a person to be held morally responsible, her biological relation-

[16] For theoretical frameworks of BWE which aim at identifying cognitive mistakes, see Köszegi and Rabin (2007, 2008), Beshears et al. (2008) and Bernheim (2016).

ship should then have the function of producing thoughts that can be assimilated to a moral continuity. Yet neither an embryo nor a skeleton seems able to produce any thought.

But this is not the main issue. Assume, for the sake of argument, that we are here only concerned with a perfectly healthy middle-age person. Assume her cerebrum is transplanted into a different living body, and that the resulting person is psychologically exactly the same as the first person (Olson 1997, pp. 43-51; DeGrazia 2005, pp. 51-54). In virtue of biological continuity, advocates of the physical view would argue that the cerebrum-less donor remains the same person, while the other cerebrum-receiver is an imposter. But as Shoemaker (2019) argues, this seems hard to believe.

There are of course some replies to this thought experiment (Olson 1997, p. 70; DeGrazia 2005, pp. 60-61) that are pointless to be discussed here. What is important to emphasise is that the physical view seems unappealing to the ethical concern of identity, and that it may provide a practical argument for tenants of personhood essentialism. In any matter, since the physical view is not endorsed by any view we are aware of in economics (except perhaps by some neuroeconomists), we will no more discuss its issues.[17] In contrast, the narrative view introduced below is certainly the one which has the most tenants in the critical literature of BWE (Sugden 2004, 2018a; Hédoin 2015; Gallois and Hédoin 2017 and Dold and Schubert 2018). Accordingly, we will spend more words discussing its philosophical issues.

5.3.3 The Narrative View

We understand how the psychological condition of identity is fundamental to ethics. Indeed, the ethical problem of identity introduced in Section 5.1 seems to ask the following question: 'what psychological characteristics are attributable to the overall individual?' It is thus not surprising that the narrative view — which is a refined version of the psychological view presented above — has the more advocates in the critical literature of BWE. This view can be expressed as follows.

> Let x be an entity that exists at time t_i and y an entity that exists at time t_j $\forall i \neq j$ and $Px \vee Py$ then $x = y$ if and only if x and y are connected by some self-told narrative relations.

Another way to say it is 'x and y can make sense of their psychological continuity'. The narrative view departs from the psychological view in the sense that it provides a meaning to the psychological relations of e.g. memories, desires and preferences. In comparison

[17] In response to Lecouteux's (2015a) point, according to which libertarian paternalism holds an implausible model of identity, Sunstein (2015) replies to the author by mentioning the possibility of considering the physical view as an alternative to Parfit's reductionist account of identity. In his words, 'Consider a competing view: in virtue of the relevant *physical* facts (for example, the same body, most importantly including the same brain), Oscar remains the same person over time.' (p. 527 — his emphasis). We however do not believe this point to be raised seriously by Sunstein (at least not in virtue of avenues of future research on justifying libertarian paternalism), considering that his initial question of which self should be attributed normative authority would otherwise be self-defeating. To be specific, if any unified view of the self is *a priori* endorsed (e.g. physicalism), there is no point assuming the multiplicity of the selves with respect to their temporality. For a defence of the physical view — often referred to as *animalism* — see Noonan (1998), Olson (2003) and Blatti and Snowdon (2016).

with the psychological view, it does not take memories, desires and preferences of one's life as merely isolated events. Instead, it *weaves them together* by giving them some form of coherence and intelligibility that they would not otherwise have had. Thus, we can see identity as a story of one's life according to the circumstances in which one's life happens (Schechtman 1996, pp. 96-99).

According to Schechtman (1996), what is more appropriate for the relation between identity and ethics is not the condition of *numerical* identity, as we formulate it in our framework by the relation $=$, but the condition of *characterisation* about one's identity. That is, the question would not be 'what are the conditions under which an individual remains one through time?' but rather 'what are the conditions under which various psychological characteristics, experiences, and actions are properly attributable to some person?' To put it differently, the question would be 'what makes the past/future states a person is specially concerned about *hers*?' We are then back to the essentialist assumption of personhood.

Like Schechtman (1996), the concept of identity endorsed by Sugden (2004, 2018a), Hédoin (2015) and Dold and Schubert (2018) seems to prioritise the characterisation condition before the numerical condition. These views may presuppose the numerical condition, but do not give an account for it. In the narrative view, what makes a psychological characteristic attributable to a person (and thus a proper part of her true self) is its 'correct' incorporation into the self-told story of her life (MacIntyre 1984, 1989; Taylor 1989; Schechtman 1996; DeGrazia 2005). Albeit it could appear that x and y are numerically different, the idea is that they can still be unified by — what we are intended to call — a phenomenological feature of their self-told narrative. Although appealing from the viewpoint of ethics, this theory of identity has however some serious flaws that we consider in turn.

First, it is left unclear why we need to tell ourselves a certain story in order to attribute ourselves unity of the events in our life, taken as a whole. As Shoemaker (2019, sec. 2.3) puts it, we may have robust psychological unity without having told ourselves any kind of story — and this story we are telling ourselves might simply be wrong (or in accordance with the vocabulary of behavioural economics, 'biased'). We might also want this narrative to be seen from a third-person standpoint, i.e. independently from the first-person standpoint. But the continuous self can constantly revise her own self-story. Another point raised by Shoemaker (2019) is that narrative unity seems to be a fuzzy condition of identity because it is left unclear that 'intelligible' actions (or choices) are those for which the individual is morally responsible. As the author argues, 'actions of children and the insane can be perfectly intelligible — even intelligible within some kind of narrative structure — without being those for which the agents are accountable' (sec. 7). In the ethical problem of identity introduced in Section 5.1, many would find unreasonable to attribute normative authority to the childhood selves, although the narrative of one's childhood may actually have the strongest structure among all one's other narratives. That is, we would be intended to think that it is not the interval of the temporal selves during childhood which is normatively relevant, but everything that happens afterwards. But tenants of narrativity would argue that we should account for *all* selves of one's life, and then weave their preferences together by some overall narrative. We however suspect many economists to reject this view because a form of 'reason' or 'ratio-

nality' seems far more appealing to characterise moral accountability than a narrative one.

Second, the narrative view endorsed in the critical literature of BWE leads to the following disturbing paradox. Recall that authors who reject the assumption of multiple selves also reject the idea that a far-sighted planner exists in virtue of her rational capacities to know what is best for her. But at the same time, they account for a narrative unity which supposes that one can — through some psychological process that is, by the way, also left unexplained — make an 'intelligible' (not to say 'rational') story by which all their choices are collected into a unified narrative. It is true that the continuous individual, as presented in the narrative view, does not presuppose to have coherent preferences at each period of time. As Sugden (2004, 2018a) puts it, the individual can have incoherent preferences and yet — we add, from a mysterious psychological ability — make 'sense' of this continuity. This would however assume that there exists a supreme self (which typically satisfies the weighting criterion seen in Section 5.1.3) that can indeed make sense and collect those incoherent preferences into a coherent (or intelligible) story. But this cannot be so, because the narrative view states that *all mental states of one's life, once gathered together meaningfully, make it the case that the self is unified*. Who this supreme 'phenomenological' self is nonetheless remains an open question. In our view, it is merely a soul or a ghost. The characterisation condition of Schechtman (1996) thus becomes unappealing to our concern because the unity of a narrative — as we have just argued — requires a unity of the self who tells such story. This ultimately presupposes strict *numerical* personal identity (MacIntyre 1984, pp. 206-208; DeGrazia 2005, p. 114). The point is that in the narrative view, one cannot be a person who has an identity unless one weaves the various experiences of one's life together into a unified story. But as Shoemaker (2019) puts it, 'the identity of that subject of experiences must be preserved across time for its experiences to be so gathered up' (sec. 2.3). This explains our commitment to the numerical identity condition.

It also explains why we consider the condition of $Px \vee Py$ instead of $Px \wedge Py$. The explanation is the following. Assume $Px \wedge Py$, and then that the identity question reduces to the question of personhood. This would mean that individuals persist only in virtue of being persons. A concept of person, broadly defined, is an individual who has the ability of being morally responsible. Thus, identity is reduced to an individual who has moral thoughts, and the question of personhood would then require an answer regarding what makes the case that an individual is a person. This account of identity would necessarily cope with the psychological view of identity, which claims that '$x = y$ if and only if x and y are connected by some given psychological relations'. By providing continuity to those psychological relations, the narrative view unifies the many experiences of one's life. But it also requires that this same individual, who can give meaning to such psychological continuity, persists through time (like e.g. an immaterial soul or a ghost), apart from the living entity at each t_i who may have incoherent preferences. Consequently, the narrative view would then be formulated as,

'$x = y$ if and only if x and y are the same unified person
who give psychological meaning to the actions of x and y'.

But 'x and y being the same person' violates our informative condition, according to which a criterion of identity cannot be trivial nor presuppose the identity it should demonstrate. For these reasons, we are inclined to reject the narrative view of personal

persistence as it presupposes the identity it is called to explain.[18]

In his theory of the individual in economics, Davis (2011) proposes what we judge to be a more compelling framework for the criterion of identity because he keeps the numerical condition. The author formulates the following two criteria of identity.

- *Individuation*. Individuals can somehow be successfully represented as distinct and independent beings.
- *Reidentification*. Individuals that have already been shown to be distinct and independent in some conception of them can be reidentified as distinct and independent in those same terms across some process of change.

As Gallois and Hédoin (2017) put it, the boundary and narrative conditions of Hédoin (2015) can be seen as respective answers to the individuation and reidentification criteria of Davis (2011) — although (as previously stated) we emphasise that narrativity should not be taken as a basic condition of identity. In our view, the reidentification criterion is a more acceptable criterion of identity since it does not presuppose the narrative nor the personhood condition. In comparison to our framework, $=$ can be understood as our individuation criterion (the fact that both persons are numerically the same at different moments of time) and Φ as our reidentification criterion (the condition which makes x and y being numerically the same individual at different moments of time).

In addition to the importance of the numerical condition over the characterisation condition, we have argued that the narrative view is philosophically problematic. So is it all what theories of personal persistence have to offer to normative economics? Before concluding, we would like to mention another unified view that may overcome some of the methodological problems of the three views previously discussed. In particular, we briefly discuss the theory of identity of Davis (2011), which is a sort of 'hybrid' theory between the narrative and the sociological view (that we now introduce).

5.3.4 The Sociological View

The sociological view (Schechtman 2014) can potentially conciliate two problems of the physical view, on the one hand, and of the psychological and narrative views, on the other hand.[19] Recall that the physical view goes too far into essentialism, and that the psychological and narrative views oppositely deny the constitution of one's identity that goes beyond one's psychology. What is nonetheless common to the biological, psychological and narrative views is that they represent identity from a *first-person* standpoint. But for each of these views, neither the social status of identity — how individuals are contextualised in their social environment — nor the story of their life told from a *third-person* standpoint is suggested. The sociological view can instead be formulated as follows.

[18] If the sceptical reader wishes to criticise our last argument, we invite her to dispute our three informative conditions of the criterion of identity over time (Section 5.2). We shall add that there are more philosophical issues related to the narrative view that we are constrained not to discuss here. For a recent assessment of the narrative view, see Olson and Witt (2019).

[19] Schechtman (2014) calls it the 'person-life view' and Shoemaker (2019) the 'anthropological view'. As we believe that 'sociological' is a terminology that better contrasts with the previous three views of identity, we prefer the latter over the two former.

Let x be an entity that exists at time t_i and y an entity that exists at time t_j
$\forall i \neq j$ and $Px \vee Py$ then $x = y$ if and only if
x and y are connected by some sociological relations.

According to Schechtman (2014 [Ch. 5]), human beings are not only characterised in virtues of their biological and psychological features, but also in virtue of their socially shaped capacities. The author considers that human beings evolve in their contextual environment — a family, a community, a nation — where these social features are essential properties of what constitutes an individual identity. That is to say, every social factor that constitutes a human being born in a given environment (her culture, habits, norms) is her ontological unit that gradually becomes responsible and concerned for its own future (Shoemaker 2019). Such responsible unit is then no different from the embryo from which she evolved, and this goes even after she dies since funerary customs preserve the identity of buried rotten skeletons. Schechtman's (2014) view of identity is familiar with the one of Davis (2011), whom the latter provides an extensive account of 'socially embedded individuals' (Ch. 3). But in contrast to Schechtman (2014), Davis (2011) precisely accounts for both narrative and sociological views of personal persistence:

> '[individuals'] self narratives about how they themselves look upon their choices trade in the language and meanings of this social discourse and cannot be understood apart from it ... From this perspective, self-narratives are both highly individualized and highly institutionalized accounts people produce to track how they see their own capability development pathways.' (p. 213)

Davis (2011) particularly criticises the model of social identity of Horst, Kirman, and Teschl (2007) for not considering individuals' preferences to be endogenously determined by their social background. Instead, he argues that those preferences have no reason to be exogenous because individuals' preferences are always changing by an 'individual-to-society' relationship he characterises through the notion of capability (Nussbaum and Sen 1993). We thus interpret the concept of identity by Davis (2011) as a 'hybrid' between the narrative and sociological views.[20]

According to us, the sociological view might be an interesting alternative to the ethical problem of identity we have so far discussed, especially when a consequent body of empirical studies in economics support the view that individual preferences are socially shaped (Akerlof and Kranton 2005; Chen and Li 2009; Benjamin et al. 2010). Insofar as the ethical identity problem of BWE has been presented, note that it was implicitly presented from a third-person standpoint, i.e. as in Sunstein's (2019) ethical concern of libertarian paternalism, which of the several selves has/have normative authority from the *social planner's* standpoint? (assuming the social planner is the ultimate judge of one's well-being). Yet the social planner is always represented as another single individual (or

[20] Some may argue that the relation between narrative and sociological views is even tighter, so tight that these account cannot be dissociated. For instance, one may argue that every narrative view should be sociological at the same time, as it is difficult to see how someone could build her own narrative in a purely introspective manner. Although this argument would deserve a longer discussion, let us accept it for the sake of argument. Even in this case, it would not follow that every sociological account is narrative, and then it would not follow that these positions cannot be dissociated. For an extensive account, see Ross (2005), who develops a 'narrative-sociological' approach where individuals progressively build their characters through strategic interactions, relying on institutions (especially language).

at best, a group of individuals), but not as the society taken as a whole.

We suspect BWE not to have implicitly assumed the sociological view of identity for the reason that it would have introduced sensitive debates about whether individuals should conform to norms, which is paradoxically already a common practice of libertarian paternalists who considers the habits of saving more and eating healthy as morally desirable (Thaler and Sunstein 2009). Also, if norms were already fully embedded into economic behaviour (e.g. it is a western norm to eat healthy, to exercise and not to smoke), then the social planner would have no role in accounting for individuals' preferences that deviate from 'good behaviour', e.g. self-control failures.

5.4 Conclusion

In the present chapter we aim to show that the assumption of the unified self in the critical literature of BWE is philosophically problematic when studied through the lens of personal persistence. We introduce the framework of personal persistence (Buonomo 2018) in order to discuss the alternative views of the unified self economists may endorse as responses to the multiple selves assumption in BWE. We emphasise that most of the unified view presented above are philosophically problematic and that the identity criterion — although when discussed from the viewpoint of ethics — is better defined by a numerical instead of a characterisation condition. In contrast with the critical literature of BWE, which mostly endorses the narrative view as an answer to the ethical issues of the multiple selves assumption, we argue that such view is philosophically problematic. Instead, economists should account for other more promising theories of personal persistence such as the sociological view.

We emphasise that considering the ethical problem of identity through the lens of the ontological approach does not mean building a theory of personal identity *per se*. Our message is that economists should not worry too much about the personhood question because the criterion of identity over time can be discussed without making any ethical claim. This is true even if their practical worries are ethical. Indeed, to provide an answer to the ethical dilemmas such as the one raised by Sunstein (2019) and Kahneman (1994), one has deep interest in first asking what makes an individual one over time, and then see how this enquiry can be informative regarding our practical ethical concerns.[21]

Many points remain of course unexplored, e.g. a detailed assessment of the sociological view or the relationship between the narrative and sociological views. Yet we uphold the crucial belief that the practical ethical appeal of a theory of personal persistence (typically the narrative view) should not divert us from our initial purpose. As we have stated it, in order to locate one's normative authority one should first aim at explaining how individuals persist through time. But it is the *result* of how individuals persist through time that has consequences on our ethical concerns about identity, and we should not see the problem upside down, i.e. 'which theory of personal persistence best fits with the

[21]There is of course the classic Hume's law objection, according to which one cannot derive ethical judgements from ontological principles. For example, even if we consensually agree that the sociological view is the 'right' one, ethical claims which would derive from this personal persistence view is another philosophical question to be solved. Although we are well aware of this potential objection, a metaethical assessment of Hume's law is (by far) outside the scope of the present chapter.

way we want to represent our idealistic picture of the morally responsible individual?'. We hope to have provided economists with a comprehensive account of the philosophical objections of the unified self, that are not only important for the future of behavioural welfare economics, but more generally to the future of normative economics.

GENERAL CONCLUSION

It is now time to conclude my journey in the relatively new but already vast area of normative behavioural economics, and as my colleague Valerio Buonomo (2019) would say, 'it is time to pay the bill' (p. 168). Since I also agree with Valerio that repetitions are boring, I will be shorter in resuming what I did and longer in drawing some lessons from the present thesis.

Summary of the Thesis

Chapter 1

The aim of this chapter is to provide a brief historical analysis of normative behavioural economics through the early discussions Kahneman and Tversky had about the normative implications of prospect theory. The usefulness of this historical analysis is to bring up new insights into the emergence of normative behavioural economics from the early 1990s to the 2000s. My main argument is that prospect theory appeared to have a substantial influential role in the development of behavioural welfare economics. As a consequence, I argue that the evolution of the heuristics-and-biases program during the 1990s is to be seen as a natural evolution in the study of well-being measurement and policy analysis rather than a strict historical break between 'positive' and 'normative' behavioural economics.

Chapter 2

In this chapter my aim is to assess from a philosophical point of view the theory of experienced utility measurement, a research program that Kahneman has recently stated to have abandoned. Because of various methodological and theoretical issues I discuss, it is argued that measuring objective happiness in terms of pleasure maximisation is flawed. Consequently, economists and policymakers have good reason to look for alternative measures of objective happiness that are not based on the maximisation of moment utilities. Instead, I suggest that economists and policymakers should rather focus on eudaimonistic conceptions of happiness that better align with the scope of public policy and with the way individuals actually perceive the notion of happiness.

Chapter 3

This chapter aims at providing a critical literature review of the reconciliation problem. As initially introduced by McQuillin and Sugden (2012), there is no consensus among scholars on how the reconciliation problem can be best tackled. My goal is thus to suggest a consensus on how the reconciliation problem can be best tackled by asking, what I judge to be, the fundamental question of this topic of research: *what is a good*

normative criterion? After presenting three important requirements that a good normative criterion should satisfy, the result is that none of the main normative criteria offered in the literature satisfy those requirements. This leads me to suggest avenues of future research on seeking ethical loci of normative economics other than *happiness*, *well-being* and *freedom*.

Chapter 4

This chapter aims at answering yet another important question of the reconciliation problem: *to whom should normative economics be addressed?* After reviewing the theoretical difficulties of the two main standpoints proposed in the literature from a social choice perspective ('view from nowhere' and 'view from somewhere'), we suggest a normative standpoint we call the 'view from *manywhere*'. In contrast to the third-person and first-person standpoint, such second-person standpoint accounts for the process of preference integration: the process by which individuals' multiple selves start with conflicting behavioural preferences and end up with their own normative preferences. The second-person standpoint implies what we call the *awareness* criterion: an individual is judged to be better off in one situation over another if her awareness set increases.

Chapter 5

In this chapter our goal is to reframe the 'multiple selves *versus* unified self' debate in normative economics within the framework of personal persistence: *what makes an individual persist from one time to another?* The reason for introducing this literature is that some authors promote the assumption of the unified self in order to palliate some of the philosophical issues related to the multiple selves assumption in behavioural welfare economics (we ourselves make such assumption in Chapter 4). Our main argument is that the unified self assumption is however no less problematic than the multiple selves assumption because it requires one to rigorously defend on ontological grounds what makes an individual persist from one time to another.

Avenues of Future Research

What then remain unexplored of the methodological and theoretical issues of normative behavioural economics? Many, and I remind my reader that the thesis only ambitions to treat a restricted number of them. There are many more paths waiting to be investigated, such as proposing an alternative normative criterion that takes an ethical locus different from happiness, well-being and freedom (as suggested in Chapter 3). With this thesis, I hope to have paved the way for a promising research agenda in normative economics, which is not only bound to behavioural economics but which is located at the intersection of identity, ethics and social choice. In what follows I briefly develop some of those avenues of future research.

On Identity

Temporal Selves and Temporal Parts

One extension of our result in Chapter 5 is to continue exploiting the literature of personal persistence by focusing on a debate that goes beyond the one between the psychological, physical, narrative and sociological views of identity. To our knowledge, all the theories of personal persistence implicitly assumed in normative economics (either 'multiple selves' or 'unified self' oriented) represent time as an exogenous variable of the persisting individual. In the introductory ethical identity problem of behavioural welfare economics (Kahneman 1994; Sunstein 2019), recall that I is composed of temporal selves who are assumed to be parts of I. Yet one difficulty, which to our knowledge has not been tackled in the literature of identity-and-economics, is the very relation between temporal selves and temporal parts. In fact, the ethical identity problem of behavioural welfare economics already makes a strong presumption about identity — that is, temporal selves are somehow coextensive to temporal parts. But instead of focusing on the 'multiple selves *versus* unified self' debate undertaken by most economists-philosophers, we believe there is more interest in focusing on the 'temporal parts *versus* not temporal parts' philosophical debate of identity.

The reason is that if we want to go ontological 'all the way down', focusing on personal persistence inevitably leads us to enter into the philosophical debate about the relationship between *parts of persons* and *time*. The literature which encapsulates this debate roughly divides in two competing theories: endurantism and perdurantism. According to endurantism, physical entities persist over time passing through time and being, strictly speaking, identical over time. That is, to say that I persists over time by enduring means that given two different times t_i and t_j, I at t_i and I at t_j is the same *entire* (or numerically identical) entity respectively at these two different times. By contrast, perdurantism says that physical entities persist by having different temporal parts at different times. Just like our common sense idea that concrete entities are composed of different spatial parts located at different regions of space, perdurantism says that they are also composed of different temporal parts located at different regions of time. Thus, according to perdurantism, concrete entities (among which living entities like individuals) are not only extended in space but also in time.

Consider the following example of endurantism. When we claim that 'I was in Reims at a workshop three days ago', it was I who saw her colleagues three days ago and who was happy to present her research. Today, I is at home. When she took the train on her way back to her home she similarly crossed time. The point is that it is not just a part of I that is at home today, with memories of her workshop in Reims. Instead, it is the *whole* I, i.e. the same individual who was at the workshop three days ago. The endurantist account of persistence sounds rather intuitive. Indeed, it is actually well in line with the way we ordinarily think about ourselves in the world. In this matter, it may support the unified view of the self. But consider now an example of perdurantism. I, from the time she arrived in Reims to the time she came back home is composed of several spatial parts, such as her head, arms, legs, and so on. What perdurantism argues is that the four-dimensional individual I has temporal parts as well, such as I-on-Sunday, I-on-Monday and I-on-Tuesday. This means that a temporal part of the four-dimensional individual I is I during an interval of time which is included in I's temporal boundaries,

namely between her departure to Reims to her arrival back at home. More generally, if a spatial part of an individual I is a part of I which is smaller than I in some spatial dimension(s), a temporal part of I is a part of I that is shorter along the temporal dimension, but which, during the relevant temporal interval, has the same spatial extent as I — i.e. it overlaps everything that is part of I during the relevant temporal interval.

A useful point we ambition to add in a further study is to put into question the fundamental presupposition about how individuals persist over time. What are the implications of a 'temporal part' approach in economics? If perdurantism is true, how should we treat temporal parts of individuals who make economic choices? As Davis (2011) puts it (and as mentioned in Chapter 5), it is a fact that intertemporal choice is largely studied in economics from both descriptive and normative aspects. But it may appear disputable to care about one's intertemporal choice if one is not the same temporal part of individual from one period to another. So what if I is composed of several parts extended through time, but that she is not, strictly speaking, the whole I at each slice of time? The point is if perdurantism is considered to be the 'right' theory of identity, it may provide an ontological defence for the multiple selves view endorsed in behavioural welfare economics. Another point of Davis (2011) is that life plans such as education, investment or health involve choices over extended selves that seem related to each other. The ontological debate between endurantism and perdurantism may then again enlighten our understanding of how selves are actually related to each other. Lastly, Davis (2011) underlines the point that individuals have a capacity to make a choice in time, which means that there is potentially 'something enduring about them apart from all their choices' (p. 6). Because we largely agree with Davis (2011) about these three points but not necessarily whether there is something *enduring* about individuals, the philosophical debate about identity in economics has the merit of being established in a more promising framework beyond the 'multiple selves *versus* unified self' debate. Instead, we believe it can be nurtured by the 'endurantism *versus* perdurantism' debate, that we keep for another study.

On Ethics

Ethically-Embedded Normative Economics

An important result of the thesis is the already widespread view in economics-and-philosophy that normative economics is ethically grounded, whatever the underlying assumption it makes about what constitutes the good life, e.g. hedonism, utilitarianism, libertarianism or virtue ethics. Consequently, an ethical theory seems necessary to normative economics if one looks for a 'convincing' normative approach. By 'convincing', I do not mean 'endorsing the most convincing ethical theory' but more modestly 'ready to be seriously defended on ethical grounds'. Examples of such normative approaches are the ones of Kahneman, Wakker, and Sarin (1997) and Sugden (2004, 2018a). Of course, I do not expect any economist to stop his interest in normative analysis if he does not have an ethical theory at hand that can provide philosophical support for his suggested normative criterion. In practice, applied economists usually have no choice but to rely on 'quick and dirty data' and then to design a preference elicitation method for a particular case of study, such as the evaluation of health states (Bleichrodt, Pinto, and Wakker 2001). One can then go along with the 'evidential view' of Hausman (2012) or with the 'clearly negative outcomes' criterion of Loewenstein and Haisley (2008) (see Section 3.2.2 in

Chapter 3). Most of the literature in behavioural welfare economics already heavily relies on these meta-criteria. The point is that, for practical purposes, it would be presumptuous to tell applied economists how they should carry out normative analysis.

But as economists-philosophers, we surely want things to be done properly. When we have the luxury option of not having to rely on 'quick and dirty' data, we want to have a normative approach that can apply to the broadest area of public policy (see the *general requirement* in Section 3.3.1, Chapter 3). In this sense, no doubt economists are on the same boat as philosophers in the journey of providing an answer to the old Socratic question of what makes the good life (or how one should live). We however still have a long way before we cross this stormy ocean. Depending on my reader's dogmatic position, the result that ethics has deep interest in being closely related to normative economics can be received as an obvious conclusion or as a very controversial claim. For some scholars, enriching normative economics with ethics may be so obvious that nothing new has been said with the present paragraph.[22] For others, it may be so wrong that I merely did not understand what normative economics is about (Gul and Pesendorfer 2007, 2008). Yet for others, this point constitutes an additional conclusion worth being taken into account, which is what I personally believe in and invite my reader to share with me. By taking the way of the methodological and theoretical issues of normative behavioural economics and by ending up with the result that ethics should be part of normative economics, this conclusion gets strengthened little by little.

Prioritise Our Ethical Concerns: Objective Criteria Before Subjective Criteria

Subjective normative criteria are associated with so many methodological and theoretical issues (see Section 3.2 in Chapter 3) that one may simply come to the radical conclusion that economists should stay away from using *subjective* normative criteria at all. In this sense, Sugden (2004, 2018a) is perhaps the first behavioural economist to take the 'objective route' with his opportunity criterion (which is independent of individuals' subjectivity). Moreover, one may fairly argue that since subjective normative criteria are, by definition, useful for assessing *individual* (or personal) states of affairs, they are of marginal help to evaluate most concerning cases at the global scale such as poverty, environmental protection, health or education. Other objective indicators of what makes the good life are matters of concern for development economics, e.g. the *Human Development Index* (Sen and Anand 1994) or the *Inclusive Wealth Index* (UNEP 2012).

The point is that there are more important things in life than limited cognitive capacities, limited attention and self-control failures. From a humanist point of view, choice situations such as saving or not, smoking or not, or eating healthy or not may not be the types of policy behavioural economists should first allocate their efforts to. We economists-philosophers and behavioural economists have perhaps something valuable to learn from the literature in development economics concerned with cases where individuals do not have the chance to access decent human living conditions. So instead of seeking subjective normative criteria that are too idiosyncratic to reflect something meaningful to the population, there is a substantial interest in focusing on *objective* normative criteria that are more likely to reach consensus on what makes individuals better off. Perhaps

[22] The reader may be provided with sufficient material in Sen (1987 [2003]). Otherwise, Broome (2009) and Mabsout (2014) provide convincing arguments.

the well-known normative criterion for that purpose is the human capability approach (Nussbaum 2000), as briefly introduced in Section 2.5, Chapter 2.

Subjective normative criteria can nonetheless also find their usefulness for most concerning issues, particularly when those issues heavily depend on individual behaviour. An illustrative example is to use the true preference criterion for environmental protection because of global warming, energy overconsumption, pollution, etc. If individuals are loss averse because they judge their well-being to be based on the reference point of what they initially possess (e.g. a fully equipped and heated house, a gasoline car and a smartphone), then the larger negative feeling of giving up those comforts compared to the smaller positive feeling of gaining those comforts may be useful information to be disclosed. Disclosing such information to the public sphere can stimulate collective discussion and deliberation on whether such consumption behaviour is desirable for the community.

In the same manner, if the adaptation effect is strong (Brickman, Coates, and Janoff-Bulman 1978) then subjective normative criteria may also provide additional information and eventually suggest paternalistic policies when the reduction of our ecological footprint and carbon emissions are no more considered to be *options* but *necessities* for our survival (see the 'clearly negative outcomes' meta-criterion in Section 3.2.2, Chapter 3). In those particular cases, *nudges* can be of good help, even if they may appear coercive to the most radical libertarians who would be reluctant to make trade-offs between freedom and survival. The point is that behavioural paternalism makes more sense when there is an externality problem (Guala and Mittone 2015). If we agree that negative externalities caused by climate change, energy overconsumption or pollution are harmful to *everyone*, a decrease in comfort and personal freedom for a greater chance of survival seems to be enough to justify paternalistic interventions.

Autonomy Before Well-Being

When uncontroversial cases such as environmental protection are set aside, lessons may however be on the side of methods to promote autonomy before well-being. I follow here the same path as my colleague Guilhem Lecouteux in his PhD thesis (2015b). The author argues that instead of judging what an erroneous choice is (which, again, is a complex enquiry for subjective criteria), we should seek to understand how individuals are able to form their own preferences (an enquiry we also suggest in Chapter 4). This is particularly necessary when autonomy defined in terms of the ability to choose one's own preference is a fundamental value that matters to us, and when we prefer to make mistakes and to learn from them rather than not being the masters of our own choices. I believe many would agree that our brain becomes more and more passive with the regular use of 'behavioural' assistance such as a GPS or self-parking cars. This is probably supported by psychology or neuroscience but for lack of empirical evidence, let us briefly consider the following hypothetical story of an individual named Ivan.

When he was an undergraduate student (back in 2010), Ivan used to be a delivery guy on a part-time job. At that time, smartphones were barely affordable. He had to use the good-old physical map of the city of Grenoble to deliver clients' orders. When he quit his job, Ivan strongly believed that he learnt the streets of his hometown faster than he would have done with the help of a GPS, even if he would have indeed delivered a bit

more quickly had he used one. Ivan now wonders whether the difference with using a GPS would have been that significant to compensate for his loss of autonomy. That is, he really doubts whether there is something more valuable than autonomy in itself, of course when nothing 'fundamentally important' is at stake. One objection to Ivan's question is that he will surely not deny that the use of a GPS would have led him to an optimal allocation of time, so that he would have either made more money (because he was paid per delivery) or the same amount of money in a shorter amount of time (and then pass on to a more pleasant activity).

The relevance of this objection depends on how 'effortful' we perceive the activity of using a physical map compared to the potential gains in money and time.[23] Is reading a plan or parking one's car not more satisfying when we know that the outcome we get from our effort results from *us*, and not from a machine that does the job for us? In 2020, we do not yet have the living conditions in Wall-E's dystopia, where the future of humankind consists in making such huge trade-off of well-being to the detriment of one's autonomy that humans are not even able to stand from their high-tech flying chairs and communicate without their virtual screens. The point is that enhancing one's autonomy appears to be much more important regarding what we are likely to gain in life, even if our well-being may be deteriorated in the short run. But autonomy and well-being can be very much compatible in the long run if we represent well-being not as an *immediate* but as *delayed* reward (hyperbolic or quasi-hyperbolic discounting caught up with us, again).

In its current state, the meaning of well-being that most behavioural economists endorse seems too conservative and is potentially influenced by the utilitarian tradition, which tends to reduce everything to a pleasure/pain calculus without distinguishing the nature of rewards (e.g. in terms of autonomy, freedom, fairness, etc.). The importance of autonomy over well-being is rising step by step in economics-and-philosophy with the advent of the competing *boost* approach to *nudge* (Grüne-Yanoff and Hertwig 2016) and with some general concerns shared by some authors. As Rizzo and Whitman (2019) put it, decision-making is like a muscle that atrophies in the long run if not used repeatedly (pp. 252-253). According to Hargreaves Heap (2017), we may want 'to feel responsible for [our] actions if those actions are to be a source of learning' (p. 257). Parfit (1984) also mentions the well-known objections to paternalism such as '[i]t is better if each of us learns from his own mistakes' and 'it is harder for others to know that these are mistakes' (p. 321). But it is yet another point to convince behavioural welfare economists that autonomy matters very much when compared to the immediate reward of using a GPS or self-parking cars.

On Social Choice

Two Literatures That Do Not (Yet) Communicate

The literature of normative behavioural economics and the one of social choice are so far, from each other, two different worlds that do not even seem to communicate. First, it does not seem that behavioural economists concerned with normative analysis take central

[23] This is without mentioning that 'effortful' seems to be an odd term to characterise such activity. Previous generations would certainly have never thought that reading a map or parking one's car will someday be considered to be effortful (possible cause: the adaptation effect, again).

issues, such as the possibility of making interpersonal comparisons of utilities or the compatibility of values, as seriously as social choice theorists.[24] This can be explained by the fact that founders of behavioural economics, who are now leading figures in normative behavioural economics (Camerer, Kahneman, Loewenstein, Thaler, among others), were mainly either psychologists or standard economists by training. Their central interests simply did not belong to the discipline of social choice.

Second, questions such as what informational basis and aggregation rules should be taken to evaluate social states of affairs inevitably lead us to the literature of social choice. Baujard (2015) makes this problem explicit for libertarian paternalism:

> 'when libertarian paternalists end up contributing to the social good, they face an aggregation issue, which is not an innocuous exercise ... Social choice theorists have long known that, without explicit normative views, [aggregating the various measures of individual well-being into one value of social welfare] is a vain attempt. Libertarian paternalists, whether they wish it or not, will again be choosing among the numerous theories of aggregation, including sum, prioritarianism, etc.' (p. 303)

These questions are specifically salient when behavioural economists retain a restricted set of values in normative analysis. Consider behavioural welfare economics. In the line of the welfarist tradition, behavioural welfare economics only takes Pareto efficiency into account. There are however other values such as liberty, autonomy, freedom, equality, fairness, etc., that are worth being taken into account in the axioms of social choice theory. The point is that if we only account for Pareto efficiency we have to ignore all the other values that matter to individuals — a well-known problem emphasised by Sen (1970 [2017]). In addition to Chapter 4, a good start to stimulate the junction between behavioural welfare economics and social choice would be to propose a version of Sen's impossibility of a Paretian Liberal applied at the individual level. Libertarian paternalism seems to be the perfect candidate to fall under this impossibility result: can we combine the Pareto-efficiency criterion with respect to individuals' liberty, knowing that individuals have non-integrated preferences?

More generally, any kind of paradox or impossibility theorem that is known in social choice theory can be transposed to the individual level if we assume an individual to be a collection of subpersonal selves with a set of preferences or strategies. Although impossibility theorems and related paradoxes are bread-and-butter issues of social choice, famous theoretical results such as the ones of Arrow (1951 [2012]) and Sen (1970 [2017]) do not seem to have bothered behavioural welfare economists so far. Consider for example the well-known Coase (1960) theorem. As fairly underlined by Hédoin (2015), 'behavioral economists have totally ignored the Coasean solution which consists in letting the agent's various selves to (interpersonally) bargain over the internalities' (p. 78). A paper concerned with the Coase solution applied to libertarian paternalism may be of good use, simply to show that the social cost problem also applies at the individual level. Indeed, it is well known in cooperative game theory that Coase theorem is not valid for more than two individuals (Gonzalez, Marciano, and Solal 2019). But

[24] Regarding the first issue, an exception is Kahneman (1999), who makes a substantial effort in providing a large amount of psychological evidence of the possibility of making interpersonal comparisons of utilities (see Section 2.3.3 and Section 2.3.4 in Chapter 2). Regarding the second issue, the problem does not apply to Sugden (2004, 2018a), who only accounts for freedom to choose as the informational basis of his normative approach — hence escaping from the conflicting relationship between well-being and freedom.

if internalities are nothing more than externalities at the intrapersonal level, one may not find it surprising that the conclusions of *social* bargaining are merely transposed to *individual* bargaining. Thus even if we set aside the ethical problem of time-inconsistent preferences in behavioural welfare economics (Chapter 5), there is still the intrapersonal bargaining problem of how to arrive to a payoff distribution with a non-empty core.[25]

Living Together

The notion of 'living together' is perhaps one of the most important goals for which policymakers should propose behavioural pubic policies. There are at least two relevant levels: the subpersonal selves of the individual (assuming they exist) and the individual members of the society. Individuals deliberate in the public sphere, but they can also revise their judgements through the process of preference formation and learning. Thus it seems important for models of endogenous preference to account for the individual level, but perhaps more importantly, also for the social level. The question is, how do individuals who form/revise their preferences *individually* then form/revise their preferences *collectively*? Such a model could guide us in the process in which policymakers can position themselves. A paper which would draw the mechanism of such deliberational process can be useful in continuation of our 'view from *manywhere*' (Chapter 4).

It is important to remind ourselves that what founds behavioural public policies are mostly social norms. For example, policymakers aim to influence employees to save for their retirement because the implicit ethical judgement conveyed (which is itself influenced by social norms) is that it is a good thing that employees save more for their retirement (Madrian and Shea 2001; Thaler and Benartzi 2004; Bernheim, Fradkin, and Popov 2015). One issue is that relying on social norms can easily tend to a slippery 'coercive' slope, where it becomes standard to think that exercising, eating five fruits/vegetables a day and not smoking is good for us. The main challenge for policymakers is then to make behavioural public policy more *transparent*, which is already a principle endorsed by Thaler and Sunstein (2009). The simple idea is that individuals can take 'enlightened' decisions if they are informed or 'aware'. Our *view from 'manywhere'* (Chapter 4) based on the deliberation process is again of good use, which accommodates well with *boost* policies (Grüne-Yanoff and Hertwig 2016).

As an example, assume the policymaker has reasonable knowledge that individuals want to consume less tobacco. A first solution would be to implement taxes: the classical tool of public economics to create incentives for not smoking. But individuals may simply be against taxes, so that another solution would be to implement *nudges*: the alternative tool of libertarian paternalism (yet at the risk of not being transparent enough). There is however another solution: behavioural public policy can be based on a deliberational consensus from which an acceptable social norm could emerge, e.g. 'it is a good thing to be prudent, to be healthy', etc. From a pragmatic viewpoint, a deliberational consensus may not be systematically unanimous (especially at the level of a whole nation). However, a large majority can be enough to justify some behavioural public policies. To palliate the problem that a majority is (by definition) not unanimous, we can supplement our

[25]The core is a technical concept in cooperative game theory. A game has a non-empty core when the set of feasible allocations among individuals cannot be improved upon by a subset (called a 'coalition') of the economy's individuals. Otherwise the core is empty. I thank Stéphane Gonzales, Philippe Solal and Kevin Techer for introducing me to this literature.

knowledge about what individuals' preferences are with theoretical explanation of how individuals form their own preferences, and with empirical surveys about what individuals think the social good is (which is the domain of descriptive ethics). The principle is that an individual cannot rationally *consent* to a policy unless this policy coerce her to make what is considered to be a good choice according to the community. Thus public intervention can be acceptable on the condition that a deliberational consensus has been reached, without waiting for a unanimous agreement to be reached (because this scenario would be very unlikely to happen).

Considering the current state of different fields, although connexions between normative behavioural economics and social choice are extremely scarce, I am confident that it will take little time until the known issues in social choice make their own way to behavioural welfare economics, and more generally to normative behavioural economics. Some connexions have already started with Hédoin's (2015, 2017) preliminary works on the ethical problems of libertarian paternalism seen from a social choice perspective, and with his book project of using the model of social choice for ethics (including ethics applied to behavioural economics) (Hédoin 2020). The author ambitions to use what he calls the 'social choice model of normative analysis' to enrich normative economics with broader notions yet unknown to this field, namely *persons*, *values* and *consent*.

We have argued in Chapter 5 that the notion of person (or personhood) is not necessarily required when questioning our practical ethical concerns, and I have previously recognised that I strongly support an extension of the set of values to be part of normative behavioural economics. The concept of *consent* is however missing in this thesis, and is a very welcome philosophical concept in order to justify behavioural public policy.[26] Hédoin (2020) takes this concept from the philosophy of Parfit (2011), which is defined as an agreement between different ethical principles of members of a society. The overall aim of Hédoin (2020) is to find an articulation between individual judgement and the social state of affairs that is to be chosen. In his view, the model of social choice allows one to accommodate the broadest range of ethical views, so that they can be compared and assessed on the basis of this useful framework.

Closing Remarks

The near future will tell us how normative behavioural economics will evolve, but I will make no prediction here. Let us see whether we will assist to the 'premature death' of behavioural welfare economics, albeit certainly for other reasons than the impossibility of making interpersonal comparisons of utilities. As for what concerns the arguments advanced in the thesis, I hope not to be considered as one of its murderers.

[26] See Baujard (2015) and Marciano (2015), who point out that libertarian paternalists mistakenly do not consider *consent* to be a serious issue for the implementation of behavioural public policy.

Bibliography

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.

Akay, A., O. Bargain, and X. Jara (2017). Back to Bentham, should we? Large-scale comparison of experienced versus decision utility. *IZA DP Working Paper N° 10907*.

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *The Quarterly Journal of Economics 115*(3), 715–753.

Akerlof, G. A. and R. E. Kranton (2005). Identity and the economics of organizations. *Journal of Economic Perspectives 19*(1), 9–32.

Akerlof, G. A. and R. E. Kranton (2010). *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-being*. Princeton University Press.

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

Alekseev, A., G. Harrison, M. Lau, and D. Ross (2019). Deciphering the noise: the welfare costs of noisy behavior. *Georgia State University Working Paper*.

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.

Angner, E. (2013). Is it possible to measure happiness? The argument from measurability. *European Journal for Philosophy of Science 3*(2), 221–240.

Aristotle (2009). *The Nicomachean Ethics*. Oxford University Press.

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

Arneson, R. J. (1990). Liberalism, distributive subjectivism, and equal opportunity for welfare. *Philosophy and Public Affairs 19*(2), 158–194.

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

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.

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

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.

Baujard, A. (2015). Beyond the consent dilemma in libertarian paternalism, a normative void. *Homo Oeconomicus 32*(2), 301–305.

Baujard, A. (2017). « L'économie du bien-être est morte. » Vive l'économie du bien-être ! ('Welfare economics is dead.' Long live to welfare economics!). In G. Campagnolo and J.-S. Gharbi (Eds.), *Philosophie Économique: Un État des Lieux*, pp. 77–129. Éditions matériologiques.

Baujard, A. and M. Gilardone (2017). Sen is not a capability theorist. *Journal of Economic Methodology 24*(1), 1–19.

Baujard, A. and M. Gilardone (2019). "Positional views" as the cornerstone of Sen's idea of justice. *GATE Working Paper*.

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

Benartzi, S. and R. H. Thaler (2002). How much is investor autonomy worth? *The Journal of Finance 57*(4), 1593–1616.

Benjamin, D. J., J. J. Choi, and A. J. Strickland (2010). Social identity and preferences. *American Economic Review 100*(4), 1913–28.

Bentham, J. (2007). *An Introduction to the Principles of Morals and Legislation*. Dover Philosophical Classics.

Berg, N. (2003). Normative behavioral economics. *The Journal of Socio-Economics 32*(4), 411–427.

Berg, N. and G. Gigerenzer (2010). As-if behavioral economics: neoclassical economics in disguise? SSRN Scholarly Paper, Social Science Research Network.

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 (2004). Addiction and cue-triggered decision processes. *American Economic Review 94*(5), 1558–1590.

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.

Beshears, J., J. J. Choi, D. Laibson, and B. C. Madrian (2008). How are preferences revealed? *Journal of Public Economics 92*(8-9), 1787–1794.

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

Bhatt, V., M. Ogaki, and Y. Yaguchi (2015). Normative behavioural economics based on unconditional love and moral virtue. *The Japanese Economic Review 66*(2), 226–246.

Bhatt, V., M. Ogaki, and Y. Yaguchi (2017). Introducing virtue ethics into normative economics for models with endogenous preferences. *Rochester Center for Economic Research Working Paper N° 600*.

Blatti, S. and P. F. Snowdon (Eds.) (2016). *Animalism: New Essays on Persons, Animals, and Identity*. Oxford University Press.

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.

Botti, S. and S. S. Iyengar (2006). The dark side of choice: when choice impairs social welfare. *Journal of Public Policy and Marketing 25*(1), 24–38.

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

Bréban, L. and M. Gilardone (2020). A missing touch of Adam Smith in Amartya Sen's account of public reasoning: the man within for the man without. *Cambridge Journal of Economics 44*(2), 257–283.

Brickman, P., D. Coates, and R. Janoff-Bulman (1978). Lottery winners and accident victims: is happiness relative? *Journal of Personality and Social Psychology 36*(8), 917–927.

Brink, D. O. (2011). Prospects for temporal neutrality. In C. Callender (Ed.), *The Oxford Handbook of Philosophy of Time*, pp. 267–286. Oxford University Press.

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.

Bruni, L. and R. Sugden (2013). Reclaiming virtue ethics for economics. *Journal of Economic Perspectives 27*(4), 141–164.

Buchanan, J. M. (1999). Natural and artifactual man. In *The Collected Works of James M. Buchanan, Volume 1: The Logical Foundations of Constitutional Liberty*, pp. 246–259. Liberty Fund.

Buonomo, V. (2018). A brief guide to personal persistence. In V. Buonomo (Ed.), *The Persistence of Persons: Studies in the Metaphysics of Personal Identity over Time*, pp. 7–18. Neunkirchen-Seelscheid: Editiones Scholasticae.

Buonomo, V. (2019). *Parts of Persons: Identity and Persistence in a Perdurantist World*. PhD thesis. Università degli studi di Milano.

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.

Carrasco, M. (2011). Hutcheson, Smith and utilitarianism. *The Review of Metaphysics 64*(3), 515–553.

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.

Chen, Y. and S. X. Li (2009). Group identity and social preferences. *American Economic Review 99*(1), 431–457.

Chernev, A., U. Böckenholt, and J. Goodman (2015). Choice overload: a conceptual review and meta-analysis. *Journal of Consumer Psychology 25*(2), 333–358.

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

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

Conee, E. (1999). Metaphysics and the morality of abortion. *Mind 108*(432), 619–646.

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

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

Darwall, S. (2006). *The Second-Person Standpoint: Morality, Respect and Accountability*. Harvard University Press.

Dasgupta, P. (2005). What do economists analyze and why: values or facts? *Economics and Philosophy 21*(2), 221–278.

Davis, J. (2018). Extending behavioral economics' methodological critique of rational choice theory. *Journal of Behavioral Economics for Policy 2*(2), 5–9.

Davis, J. B. (2003). *The Theory of the Individual in Economics: Identity and Value*. Routledge.

Davis, J. B. (2011). *Individuals and Identity in Economics*. Cambridge University Press.

Davis, J. B. (2016). Economists' odd stand on the positive-normative distinction: a behavioral economics view. In G. F. DeMartino and D. McCloskey (Eds.), *The Oxford Handbook of Professional Economic Ethics*, pp. 199–218. Oxford University Press.

DeGrazia, D. (2005). *Human Identity and Bioethics*. Cambridge University Press.

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

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

Do, A., A. Rupert, and G. Wolford (2008). Evaluations of pleasurable experiences: the peak-end rule. *Psychonomic Bulletin & Review 15*, 96–98.

Dolan, P. and D. Kahneman (2008). Interpretations of utility and their implications for the valuation of health. *The Economic Journal 118*(525), 215–234.

Dold, M. F. (2017). *Non-Standard Preferences, Welfare, and Public Policy: Methodological and Normative Implications of Behavioral Economics*. PhD thesis. Albert-Ludwigs-Universität Freiburg.

Dold, M. F. (2018). Back to Buchanan? Explorations of welfare and subjectivism in behavioral economics. *Journal of Economic Methodology 25*(2), 160–178.

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

Dold, M. F. and A. Stanton (2020). I choose for myself, therefore I am: the contours of existentialist welfare economics (forthcoming). *Erasmus Journal for Philosophy and Economics*.

Edgeworth, F. Y. (1881). *Mathematical Psychics*. C. Kegan Paul & Co.

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

Elster, J. (1998). *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge University Press.

Elster, J. (2000). *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*. Cambridge University Press.

Elster, J. (2016). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge University Press.

Epicurus (1994). *Epicure : Lettres, Maximes, Sentences (Epicurus: Letters, Maxims, Sentences)*. Le Livre de Poche.

Ferey, S. (2011). Paternalisme libéral et pluralité du moi (Libertarian paternalism and multiple selves). *Revue Économique 62*(4), 737–750.

Fine, B. (1995). On the relationship between true preference and actual choice. *Social Choice and Welfare 12*, 353–361.

Fleurbaey, M. and P. Hammond (2004). Interpersonally comparable utility. In S. Barbera, P. Hammond, and C. Seidl (Eds.), *Handbook of Utility Theory: Volume 2 Extensions*, pp. 1179–1285. Springer.

Fredrickson, B. L. and D. Kahneman (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology 65*, 45–55.

Frey, B. S. and A. Stutzer (2002). What can economists learn from happiness research? *Journal of Economic Literature 40*(2), 402–435.

Fumagalli, R. (2013). The futile search for true utility. *Economics and Philosophy 29*(3), 325–347.

Gallois, F. and C. Hédoin (2017). From identity to agency in positive and normative economics. In *Forum for Social Economics*, pp. 1–17. Taylor & Francis.

Gauthier, D. (1986). *Morals by Agreement*. Clarendon Press.

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

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky. *Psychological Review 103*(3), 592–596.

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology 6*, 361–383.

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

Gigerenzer, G. and R. Selten (Eds.) (2001). *Bounded Rationality: The Adaptive Toolbox*. MIT Press.

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

Goodin, R. (1992). Laundering Preferences. In J. Elster and A. Hylland (Eds.), *Foundations of Social Choice Theory* (reprinted ed.)., pp. 75–101. Cambridge University Press.

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. (2016). Why behavioural policy needs mechanistic evidence. *Economics and Philosophy 32*(3), 463–483.

Grüne-Yanoff, T. (2018). Boosts vs. nudges from a welfarist perspective. *Revue d'Économie Politique 128*, 209–224.

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

Guala, F. and L. Mittone (2015). A political justification of nudging. *Review of Philosophy and Psychology 6*(3), 385–395.

Gul, F. and W. Pesendorfer (2001). Temptation and self-control. *Econometrica 69*(6), 1403–1435.

Gul, F. and W. Pesendorfer (2007). Welfare without happiness. *American Economic Review 97*(2), 471–476.

Gul, F. and W. Pesendorfer (2008). The case for mindless economics. In A. Caplin and A. Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook*, pp. 3–39. Oxford University Press.

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

Hands, D. W. (2010). Economics, psychology and the history of consumer choice theory. *Cambridge Journal of Economics 34*(4), 633–648.

Hands, D. W. (2012). The positive-normative dichotomy and economics. In U. Mäki (Ed.), *Philosophy of Economics*, pp. 219–239. North Holland.

Hands, D. W. (2014). Normative ecological rationality: normative rationality in the fast-and-frugal-heuristics research program. *Journal of Economic Methodology 21*(4), 396–410.

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

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

Harrison, G. W. (2019). The behavioral welfare economics of insurance. *The Geneva Risk and Insurance Review 44*(2), 137–175.

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

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*.

Harsanyi, J. C. (1977). Rule utilitarianism and decision theory. *Erkenntnis 11*(1), 25–53.

Hausman, D. M. (2008). Mindless or mindful economics: a methodological evaluation. In A. Caplin and A. Schotter (Eds.), *The Foundations of Positive and Normative Economics*, pp. 126–151. Oxford University Press.

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

Hausman, D. M. (2015). *Valuing Health: Well-Being, Freedom, and Suffering*. Oxford University Press.

Hausman, D. M. (2016). On the Econ within. *Journal of Economic Methodology 23*(1), 26–32.

Hausman, D. M. (2018). The bond between positive and normative economics. *Revue d'Économie Politique 128*(2), 191–208.

Hausman, D. M., M. McPherson, and D. Satz (2016). *Economic Analysis, Moral Philosophy, and Public Policy* (third ed.). Cambridge University Press.

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

Hédoin, C. (2017). Normative economics and paternalism: the problem with the preference-satisfaction account of welfare. *Constitutional Political Economy 28*(3), 286–310.

Hédoin, C. (2020). *Persons, Values and Consent: From Social Choice to Social Philosophy*. Unpublished Book Project.

Heidhues, P., J. Johnen, and B. Köszegi (2020). Browsing versus studying: a pro-market case for regulation (forthcoming). *Review of Economic Studies*.

Herrnstein, R. J., G. Loewenstein, D. Prelec, and W. Vaughan Jr. (1993). Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making 6*(3), 149–185.

Hershenov, D. (2005). Do dead bodies pose a problem for biological approaches to personal identity? *Mind 114*(453), 31–59.

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

Horst, U., A. Kirman, and M. Teschl (2007). Changing identity: the emergence of social groups. *Princeton, NJ: Institute for Advanced Study, School of Social Science, Economics Working Papers*.

Hursthouse, R. (2016). Virtue ethics. *The Stanford Encyclopedia of Philosophy*.

Hutchinson, J. M. C. (2005). Is more choice always desirable? Evidence and arguments from leks, food selection, and environmental enrichment. *Biological Reviews 80*, 73–92.

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.

Iyengar, S. S. (2010). *The Art of Choosing*. Hachette Digital.

Iyengar, S. S. and M. R. Lepper (2000). When choice is demotivating: can one desire too much of a good thing? *Journal of Personality and Social Psychology 79*(6), 995–1006.

Jevons, W. S. (1905). *Essays on Economics*. Macmillan.

Jullien, D. (2016). All frames created equal are not identical: on the structure of Kahneman and Tversky's framing effects. *Œconomia. History, Methodology, Philosophy 6*(2), 265–291.

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.

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. (2000). Experienced utility and objective happiness: a moment-based approach. In D. Kahneman and A. Tversky (Eds.), *Choices, Values, and Frames*, pp. 673–692. Cambridge University Press.

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

Kahneman, D., B. L. Fredrickson, C. A. Schreiber, and D. A. Redelmeier (1993). When more pain is preferred to less: adding a better end. *Psychological Science 4*(6), 401–405.

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., J. L. Knetsch, and R. H. Thaler (1991). Anomalies: the endowment effect, loss aversion, and status quo bias. *The Journal of Economic Perspectives 5*(1), 193–206.

Kahneman, D. and A. B. Krueger (2006). Developments in the measurement of subjective well-being. *Journal of Economic Perspectives 20*(1), 3–24.

Kahneman, D., A. B. Krueger, D. A. Schkade, N. Schwarz, and A. A. Stone (2004). A survey method for characterizing daily life experience: the day reconstruction method. *Science 306*(5702), 1776–1780.

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 J. S. Snell (1992). Predicting a changing taste: do people know what they will like? *Journal of Behavioral Decision Making 5*(3), 187–200.

Kahneman, D. and R. Sugden (2005). Experienced utility as a standard of policy evaluation. *Environmental and Resource Economics 32*(1), 161–181.

Kahneman, D. and R. H. Thaler (2006). Anomalies: utility maximization and experienced utility. *Journal of Economic Perspectives 20*(1), 221–234.

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 A. Tversky (Eds.) (2000). *Choices, Values, and Frames*. Cambridge University Press.

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.

Karlsson, N., G. Loewenstein, and J. McCafferty (2004). The economics of meaning. *Nordic Journal of Political Economy 30*, 61–75.

Kemp, S., C. D. B. Burt, and L. Furneaux (2008). A test of the peak-end rule with extended autobiographical events. *Memory & Cognition 36*(1), 132–138.

Kincaid, H., J. Dupré, and A. Wylie (2007). *Value-Free Science?: Ideals and Illusions*. Oxford University Press.

Kirman, A. and M. Teschl (2004). On the emergence of economic identity. *Revue de Philosophie Économique 9*(1), 59–86.

Korsgaard, C. M. (1989). Personal identity and the unity of agency: a Kantian response to Parfit. *Philosophy & Public Affairs 18*(2), 101–132.

Köszegi, B. and M. Rabin (2007). Mistakes in choice-based welfare analysis. *American Economic Review 97*(2), 477–481.

Köszegi, B. and M. Rabin (2008). Revealed mistakes and revealed preferences. In A. Caplin and A. Schotter (Eds.), *The Foundations of Positive and Normative Economics*, pp. 193–209. Oxford University Press.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics 112*(2), 443–478.

Larrouy, L. and G. Lecouteux (2018). Choosing in a large world: the role of focal points as a mindshaping device. *GREDEG Working Paper*.

Larson, R. and M. Csikszentmihalyi (1983). The experience sampling method. *New Directions for Methodology of Social and Behavioral Science 15*, 41–56.

Layard, R. (2011). *Happiness: Lessons from a New Science* (second ed.). Penguin Books.

Lecouteux, G. (2015a). In search of lost nudges. *Review of Philosophy and Psychology 6*(3), 397–408.

Lecouteux, G. (2015b). *Reconciling Normative and Behavioural Economics*. PhD thesis. École Polytechnique.

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. (1987). Anticipation and the valuation of delayed consumption. *The Economic Journal 97*(387), 666–684.

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

Loewenstein, G. (1999). Because it is there: the challenge of mountaineering ... for utility theory. *Kyklos 52*(3), 315–343.

Loewenstein, G. and E. Haisley (2008). The economist as therapist: methodological ramifications of "light" paternalism. In A. Caplin and A. Schotter (Eds.), *The Foundations of Positive and Normative Economics*, pp. 210–245. Oxford University Press.

Loewenstein, G. and R. H. Thaler (1989). Anomalies: intertemporal choice. *Journal of Economic Perspectives 3*(4), 181–193.

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

Lowe, E. J. (2012). The probable simplicity of personal identity. In G. Gasser and M. Stefan (Eds.), *Personal Identity: Complex or Simple?*, pp. 137–155. Cambridge University Press.

Mabsout, R. (2014). Bringing ethics back to welfare economics. *Review of Social Economy 72*(1), 1–27.

MacIntyre, A. (1984). *After Virtue*. University of Notre Dame Press.

MacIntyre, A. (1989). The virtues, the unity of a human life and the concept of a tradition. In S. Hauerwas and L. G. Jones (Eds.), *Why Narrative?* W.B. Eerdmans.

Madrian, B. C. (2014). Applying insights from behavioral economics to policy design. *Annual Review of Economics 6*(1), 663–688.

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.

Mah, E. Y. and D. M. Bernstein (2019). No peak-end rule for simple positive experiences observed in children and adults. *Journal of Applied Research in Memory and Cognition 8*(3), 337–346.

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

March, J. G. (1978). Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics 9*, 587–608.

Marciano, A. (2015). The consent dilemma in libertarian paternalism. *Homo Oeconomicus 32*(2), 287–291.

Martin, R. (1998). *Self-Concern: An Experiential Approach to What Matters in Survival*. Cambridge University Press.

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2012). Revealed attention. *American Economic Review 102*(5), 2183–2205.

McCloskey, D. N. (2019). *Economical Writing* (third ed.). Chicago University Press.

McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford University Press.

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

Mill, J. S. (1972). *On Liberty*. Dent.

Mitchell, G. (2005). Libertarian paternalism is an oxymoron. *Northwestern University Law Review 99*(3).

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

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

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

Nagatsu, M. (2015b). Social nudges: their mechanisms and justification. *Review of Philosophy and Psychology 6*(3), 481–494.

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

Noonan, H. W. (1998). Animalism versus Lockeanism: a current controversy. *The Philosophical Quarterly 48*(192), 302–318.

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

Nozick, R. (1981). *Philosophical Explanations*. Harvard University Press.

Nussbaum, M. (2000). *Women and Human Development: The Capabilities Approach*. Cambridge University Press.

Nussbaum, M. C. and A. Sen (Eds.) (1993). *The Quality of Life*. Oxford University Press.

O'Donoghue, T. and M. Rabin (1999). Doing it now or later. *American Economic Review 89*(1), 103–124.

Ogaki, M. and S. C. Tanaka (2017). *Behavioral Economics: Toward a New Economics by Integration with Traditional Economics*. Springer.

Oliver, A. (2017). Distinguishing between experienced utility and remembered utility. *Public Health Ethics 10*(2), 122–128.

Olson, E. (1997). *The Human Animal: Personal Identity without Psychology*. Oxford University Press.

Olson, E. T. (2003). An argument for animalism. In R. Martin and J. Barresi (Eds.), *Personal Identity*, pp. 318–334. Blackwell.

Olson, E. T. and K. Witt (2019). Narrative and persistence. *Canadian Journal of Philosophy 49*(3), 419–434.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Parfit, D. (2011). *On What Matters*, Volume 1. Oxford University Press.

Pattanaik, P. K. and Y. Xu (1990). On ranking opportunity sets in terms of freedom of choice. *Recherches Économiques de Louvain (Louvain Economic Review) 56*(3-4), 383–390.

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.

Plott, C. R. (1996). Rational individual behavior in markets and social choice processes: the discovered preference hypothesis. In K. J. Arrow, E. Colombatto, M. Perlaman, and C. Schmidt (Eds.), *The Rational Foundations of Economic Behaviour*, pp. 225–250. McMillian.

Putnam, H. (2002). *The Collapse of the Fact/Value Dichotomy and Other Essays*. Harvard University Press.

Putnam, H. and V. Walsh (Eds.) (2011). *The End of Value-Free Economics*. Routledge.

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

Railton, P. (2003). *Facts, Values, and Norms: Essays toward a Morality of Consequence*. Cambridge University Press.

Read, D. (2007). Experienced utility: utility theory from Jeremy Bentham to Daniel Kahneman. *Thinking & Reasoning 13*(1), 45–61.

Rebonato, R. (2012). *Taking Liberties: A Critical Examination of Libertarian Paternalism*. Palgrave Macmillan UK.

Rebonato, R. (2014). A critical assessment of libertarian paternalism. *Journal of Consumer Policy 37*(3), 357–396.

Redelmeier, D. A. and D. Kahneman (1996). Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain 66*(1), 3–8.

Redelmeier, D. A., J. Katz, and D. Kahneman (2003). Memories of colonoscopy: a randomized trial. *Pain 104*(1), 187–194.

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

Reisch, L. A., C. R. Sunstein, and W. Gwozdz (2017). Beyond carrots and sticks: Europeans support health nudges. *Food Policy 69*, 1–10.

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

Rizzo, M. J. and D. G. Whitman (2018). Rationality as a process. *Review of Behavioral Economics 5*(3-4), 201–219.

Rizzo, M. J. and D. G. Whitman (2019). *Escaping Paternalism*. Cambridge University Press.

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

Ross, D. (2014). *Philosophy of Economics*. Palgrave Macmillan.

Rovane, C. (1998). *The Bounds of Agency*. Princeton University Press.

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

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

Samuelson, P. (1937). A note on measurement of utility. *Review of Economic Studies 4*(2), 155–161.

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

Savage, L. J. (1972). *The Foundations of Statistics*. Courier Corporation.

Schechtman, M. (1996). *The Constitution of Selves*. Cornell University Press.

Schechtman, M. (2014). *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*. Oxford University Press.

Scheibehenne, B. (2008). *The Effect of Having Too Much Choice*. PhD thesis. Humboldt-Universität zu Berlin.

Scheibehenne, B., R. Greifeneder, and P. M. Todd (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research 37*(3), 409–425.

Schkade, D. A. and D. Kahneman (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science 9*(5), 340–346.

Schreiber, C. A. and D. Kahneman (2000). Determinants of the remembered utility of aversive sounds. *Journal of Experimental Psychology: General 129*(1), 27–42.

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

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

Scoccia, D. (2019). Paternalisms and nudges. *Economics and Philosophy 35*(1), 79–102.

Sen, A. (1977). On weights and measures: informational constraints in social welfare analysis. *Econometrica 45*(7), 1539–1572.

Sen, A. (1985). *Commodities and Capabilities*. North Holland.

Sen, A. (1991). Welfare, preference and freedom. *Journal of Econometrics 50*(1-2), 15–29.

Sen, A. (1993). Positional objectivity. *Philosophy and Public Affairs 22*(2), 126–145.

Sen, A. (2003). *On Ethics and Economics* (reprinted ed.). Blackwell.

Sen, A. (2006). Reason, freedom and well-being. *Utilitas 18*(1), 80–96.

Sen, A. (2009). *The Idea of Justice*. Harvard University Press.

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

Sen, A. and S. Anand (1994). Human development index: methodology and measurement. *Human Development Report Office: Occasional Papers*.

Shoemaker, D. (2019). Personal identity and ethics. *The Stanford Encyclopedia of Philosophy*.

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

Smith, A. (2010). *Theory of Moral Sentiments*. Penguin Classics.

Smith, D. M., R. L. Sherriff, L. J. Damschroder, G. Loewenstein, and P. A. Ubel (2006). Misremembering colostomies? Former patients give lower utility ratings than do current patients. *Health Psychology 25*(6), 688–695.

Smith, V. L. (2003). Constructivist and ecological rationality in economics. *American Economic Review 93*(3), 465–508.

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

Sudgen, R. (1998). The metric of opportunity. *Economics and Philosophy 14*(2), 307–337.

Sugden, R. (2003). Opportunity as a space for individuality: its value and the impossibility of measuring it. *Ethics 113*(4), 783–809.

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

Sugden, R. (2006). What we desire, what we have reason to desire, whatever we might desire: Mill and Sen on the value of opportunity. *Utilitas 18*(1), 33–51.

Sugden, R. (2007). The value of opportunities over time when preferences are unstable. *Social Choice and Welfare 29*(4), 665–682.

Sugden, R. (2008). Why incoherent preferences do not justify paternalism. *Constitutional Political Economy 19*(3), 226–248.

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

Sugden, R. (2013). The behavioural economist and the social planner: to whom should behavioural welfare economics be addressed? *Inquiry 56*(5), 519–538.

Sugden, R. (2015). Looking for a psychology for the inner rational agent. *Social Theory and Practice 41*(4), 579–598.

Sugden, R. (2017a). Characterising competitive equilibrium in terms of opportunity. *Social Choice and Welfare 48*(3), 487–503.

Sugden, R. (2017b). Do people really want to be nudged towards healthy lifestyles? *International Review of Economics 64*(2), 113–123.

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

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

Sunstein, C. R. (2015). Nudges, agency, and abstraction: a reply to critics. *Review of Philosophy and Psychology 6*(3), 511–529.

Sunstein, C. R. (2019). *On Freedom*. Princeton University Press.

Sunstein, C. R., L. A. Reisch, and M. Kaiser (2019). Trusting nudges? Lessons from an international survey. *Journal of European Public Policy 26*(10), 1417–1443.

Swinburne, R. (1984). Personal identity: the dualist theory. In S. Sydney and R. Swinburne (Eds.), *Personal Identity*, pp. 3–66. Blackwell.

Taylor, C. (1989). *Sources of the Self: The Making of the Modern Identity*. Harvard University Press.

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 (1991). Loss aversion in riskless choice: a reference-dependent model. *The Quarterly Journal of Economics 106*(4), 1039–1061.

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

Tversky, A. and R. H. Thaler (1990). Anomalies: preference reversals. *Journal of Economic Perspectives 4*(2), 201–211.

UNEP (2012). *Inclusive Wealth Report 2012: Measuring Progress toward Sustainability*. Cambridge University Press.

Varey, C. A. and D. Kahneman (1992). Experiences extended across time: evaluation of moments and episodes. *Journal of Behavioral Decision Making 5*(3), 169–185.

Veenhoven, R. (2000). Freedom and happiness: a comparative study in forty-four nations in the early 1990s. In E. Diener and E. M. Suh (Eds.), *Culture and Subjective Well-Being*, pp. 257–88. MIT Press.

Welch, B. and D. Hausman (2010). To nudge or not to nudge? *Journal of Political Philosophy 18*(1), 123–136.

Whitman, D. G. and M. J. Rizzo (2015). The problematic welfare standards of behavioral paternalism. *Review of Philosophy and Psychology 6*(3), 409–425.

Abstract

Abstract in English

This thesis is a collection of five chapters which tackle and aim to solve various methodological and theoretical issues associated with normative behavioural economics. The first chapter proposes a historical reconstruction of normative behavioural economics. It is shown that the founders of prospect theory already had an early interest in the normative implications of their theory, which had a substantial influence on the methodology of behavioural welfare economics. The second chapter is a philosophical assessment of the theory of experienced utility measurement. After showing that the experienced utility criterion suffers from many methodological and theoretical problems, I suggest an alternative approach of objective happiness that aligns better with the scope of public policy and with the way individuals actually perceive the notion of objective happiness. The third chapter proposes a literature review of the 'problem of reconciling' normative and behavioural economics. I suggest a consensus on how the 'reconciliation problem' can be best tackled by proposing a simple framework by which economists could consensually agree about what a 'good' normative criterion is. The result is however that none of the main normative criteria offered in the literature satisfy all requirements of the proposed framework. In the fourth chapter, we propose an alternative form of normative economics that accounts for context-dependent preferences. Our approach differs from other approaches offered in the literature in the sense that it focuses on the process by which individuals' multiple selves start with conflicting preferences and end up with their own preferences (an approach we label 'view from *manywhere*'). In the fifth and last chapter, we introduce the ontological framework of personal persistence in normative economics in order to discuss some ethical concerns of time-inconsistent preferences. The overall result of the present thesis is that albeit normative behavioural economics rapidly flourished over the last few years in public policy, this domain of research still needs to address a consequent number of methodological and theoretical issues before it can be considered as a promising field to be applied in public decision-making. Normative behavioural economics must specially face two important problems, which result from those already studied in this thesis. First, the ethical issues related to time-inconsistent preferences require the improvement of our ontological understanding of individual identity. Second, the theoretical problems of normative behavioural economics require to be assessed by the tools of social choice: a rigorous framework that would allow us to clarify in formal language several theoretical objections listed in the critical literature.

Keywords: *choice — cognitive biases — ethics — identity — preference — public policy — rationality — well-being*

JEL codes: B41, D60, D90, I31

Résumé en Français

Cette thèse est un recueil de cinq chapitres qui abordent et visent à résoudre divers problèmes méthodologiques et théoriques associés à l'économie comportementale normative. Le premier chapitre propose une reconstruction historique de l'économie comportementale normative. Il est montré que les fondateurs de la prospect theory s'intéressaient déjà aux implications normatives de leur théorie, ce qui a eu une influence substantielle sur la méthodologie de l'économie comportementale du bien-être. Le deuxième chapitre est une évaluation philosophique de la théorie de la mesure d'utilité expérimentée. Après avoir montré que le critère d'utilité expérimentée souffre de nombreux problèmes méthodologiques et théoriques, je propose une approche alternative du bonheur objectif mieux alignée avec la portée des politiques publiques et avec la manière dont les individus perçoivent réellement la notion de bonheur objectif. Le troisième chapitre propose une revue de la littérature du « problème de réconciliation » entre économie normative et économie comportementale. Je suggère un consensus sur la meilleure façon de traiter le « problème de réconciliation » en proposant un cadre simple sur lequel les économistes pourraient s'entendre sur ce qu'est un « bon » critère normatif. Le résultat est qu'aucun des principaux critères normatifs proposés dans la littérature ne satisfait cependant à toutes les exigences du cadre proposé. Dans le quatrième chapitre, nous proposons une forme alternative d'économie normative qui tient compte des préférences dépendantes du contexte. Notre approche diffère des autres approches proposées dans la littérature dans la mesure où elle se concentre sur le processus par lequel les moi multiples de l'individu commencent par des préférences conflictuelles et aboutissent à leurs propres préférences (une approche que nous appelons « vue de nombreuses positions »). Dans le cinquième et dernier chapitre, nous introduisons le cadre ontologique de la persistance personnelle en économie normative afin de discuter certains problèmes éthiques de préférences incohérentes dans le temps. Le résultat général de cette thèse est que malgré la prolifération rapide de l'économie comportementale normative dans le domaine de la politique publique, l'économie comportementale normative doit encore pallier un nombre conséquent de problèmes méthodologiques et théoriques avant de pouvoir s'affirmer en tant que champ prometteur dans la décision publique. Ce champ de recherche doit notamment faire face à deux problèmes importants qui résultent de ceux déjà étudiés dans la présente thèse. Premièrement, les problèmes éthiques liés aux changements de préférences dans le temps requièrent d'améliorer la compréhension ontologique de l'identité individuelle. Secondement, les problèmes théoriques de l'économie comportementale normative nécessitent d'être évalués par les outils du choix social : un cadre rigoureux qui nous permettrait de clarifier dans un langage formel plusieurs objections théoriques répertoriées dans la littérature critique.

Mots-clés : *choix — biais cognitifs — éthique — identité — préférence — politique publique — rationalité — bien-être*

Codes JEL : B41, D60, D90, I31