A methodological perspective on behavioral economics and the role of language in economic rationality

Dorian Jullien

Thèse de doctorat

L'économie comportementale et le rôle du langage dans la rationalité économique : une perspective méthodologique.

A methodological perspective on behavioral economics and the role of language in economic rationality

Présentée en vue de l'obtention du
grade de docteur de l'Université de Nice Sophia-Antipolis

École doctorale N°513 : *ED-DESPEG*

Discipline : ***Sciences Économiques***

Soutenue le : 8 juin 2016

Par

Dorian JULLIEN

JURY

Rapporteurs : Annie COT, Professeur, Paris 1 Panthéon-Sorbonne
Wade HANDS, Professeur, University of Puget Sound

Directeur de thèse : Richard ARENA, Professeur, Université de Nice Sophia-Antipolis

Membres : Matthias KLAES, Professeur, University of Dundee
Guillaume HOLLARD, Directeur de recherche au CNRS, Ecole Polytechnique
Agnès FESTRE, Professeur, Université de Nice Sophia-Antipolis Membres :
Christian HUDELOT, Directeur de recherche au CNRS, Université de Nice Sophia-Antipolis

I believe that I can best make the relation of my ideography to ordinary language clear if I compare it to that which the microscope has to the eye. Because of the range of its possible uses and the versatility with which it can adapt to the most diverse circumstances, the eye is far superior to the microscope. Considered as an optical instrument, to be sure, it exhibits many imperfections, which ordinarily remain unnoticed only on account of its intimate connection with our mental life. But, as soon as scientific goals demand great sharpness of resolution, the eye proves to be insufficient . The microscope, on the other hand, is perfectly suited to precisely such goals, but that is just why it is useless for all others.

> Gottlob Frege, 1878, in the Preface to *Begriffsschrift* (translation by Stefan Bauer-Mengelberg)

Charlie: Right. Let's slash his tires.
Mac: Well, not that, though, because then he can't leave. That doesn't make any sense.
Charlie: Well, you start putting plans under microscopes, nothing's gonna make sense, all right?
Mac: Lots of things make sense! Slashing someone's tires so they leave makes no sense, Charlie.
Charlie: Are you gonna put everything I say under a microscope, bud?
Dennis: It's a stupid idea, Charlie.
Charlie: I'm sorry, do you have a better plan?

> Charlie Day, Rob McElhenney and Glenn Howerton, 2009, in the opening scene to "The Great Recession": E03S05 of *It's Always Sunny in Philadelphia* (written by David Hornsby, Chadd Gindin, Rob McElhenney and Glenn Howerton)

Contents

List of Figures

List of Tables

General Introduction

"Ordinary language and ordinary thought have been the victims of neglect in economic science in general. Their potential for the explanation of the most economic of phenomena may, however, be great." (Bacharach 1990, p.368)

"[E]conomists seem to have largely ignored language. This is an unfortunate state of affairs. The world people live in is a world of words, not functions, and many real phenomena might be more easily analyzed if we take this into account" (Lipman 2003, p.75)

"[M]ost information sharing is not done through Spence-style signaling, through the price system, nor through carefully crafted Hurwicz-style incentive compatible mechanisms: it is done through ordinary, informal talk." (Farell and Rabin 1996, p.104)

Over the last three decades, the standard economic analysis of individual behaviors has been subjected to the relentless criticisms of an approach known as 'behavioral economics'. An ever growing set of empirical regularities is presented by behavioral economists as contradicting the predictions of theoretical models used in standard economics. These criticisms are accompanied by various methodological considerations, notably on the lack of interdisciplinary relations between economics and psychology, or about the role of the positive/normative distinction within

models of individual behaviors in economics. Over the last decade, standard economists have increasingly responded to the these criticisms, albeit not in a homogenous way. Some standard economists have granted the soundness of *parts* of behavioral economists' criticisms; other ones have granted the soundness of *all* of them to the point of now practicing behavioral economics and calling themselves 'behavioral economists'. Yet some others have criticized back, arguing notably that behavioral economists misunderstood the standard approach. In short, the frontier between behavioral and standard economics is today fuzzier than it was thirty years ago, as are the connections between the theoretical, empirical and methodological issues underlying it. This fuzzy frontier constituted by the behavioral *versus* standard economics debates is the object of the present dissertation. The main intended contribution is to try to provide a better understanding of the theoretical and methodological ongoing transformations to the economic analysis of individual behaviors impulsed by behavioral economics.

One of the causes of the fuzziness of the frontier is that behavioral economics is an approach that uses insights from the discipline of psychology to *modify* the standard models of individual behaviors in economics. In other words, the traditional *homo œconomicus* – always making correct reasoning, never regretting past choices and guided only by self-interest – is modified to take a more human face, or so behavioral economists argue. This argument is systematized by the father of behavioral economics, Richard Thaler (2015; 2016), who contrasts the species of 'Econs' represented in standard economics' models with the species of 'Humans' represented in his approach. In behavioral economics, he argues, economic agents are represented as *sometimes* making mistaken choices and reasoning, as well as *not always* behaving regarding their own self-interest. In other words, behavioral economics is presented as a generalization of standard economics: Econs never behave like Humans but Humans sometimes behave like Econs. One of the reasons why Thaler and other behavioral economists do not completely criticize standard economics is that they need it as a normative benchmark to characterize what counts as *rational* behaviors. Hence, psychology is used in behavioral economics for the purpose of modeling the departures between how economic agents *are* behaving and how they *ought* to behave rationally. The goal of this dissertation is to try to show that a non-trivial aspect of what makes economic agents Humans remains largely unexplored in behavioral economics and that exploring it can actually shed some useful lights on the standard *versus* behavioral economics debates. This

aspect is quite simply that *we talk*. That is, we use language in ways other species, including Econs, do not. How does the uses of language play a role in economic rationality that can help us better understand the standard *versus* behavioral economics debates?

This dissertation argues that the uses of language do not play one but two roles in economic rationality. On the one hand, economists use language to discuss, dispute and theorize economic rationality. On the other hand, economic agents use language to make economic choices, or, more importantly, to construct decision problems in which other agents make economic choices. As Ariel Rubinstein puts it:

> "*economic agents* are human beings for whom language is a central tool in the process of making decisions and forming judgments. And [...] the other important "players" in Economic Theory – namely ourselves, the *economic theorists* – use formal models but these are not simply mathematical models; their significance derives from their interpretation, which is expressed using daily language." (2000 pp.3-4, my emphases)

In other words, the role of language in economic rationality is twofold: the behaviors of economic agents are partly constituted of language uses, and so are the professional activities of economists, especially theoreticians. Two sets of classical distinctions used among linguists and philosophers of language are useful to clarify further how the role of language in economic rationality is studied in this dissertation. The first one, made by Ferdinand de Saussure (1916, chap.3), is between language (*langage*), a language (*langue*) and speech (*parole*). Roughly, 'language' refers to the overarching set of all *communicative devices* (e.g., symbols, gestures, phonemes, etc.), a particular conventional articulation of which constitutes 'a language' used in a *community* (e.g., English, French, local dialects etc.), which is *individually* used and instantiated by 'speech' (e.g., written or spoken words and sentences). The second set of classical distinctions, made by Charles W. Morris in 1938, is between three types of relations among the symbols constitutive of a language (see Posner 1987, sect.1): syntax, semantics and pragmatics. Roughly, syntax designates the relations *between symbols*, semantics designates the relations *between symbols and other entities* (e.g., 'objects', 'things', 'other people' etc.) that are neither symbols nor the human who is using them and pragmatics designates the relations *between symbols and the human who is using them* (e.g., what are his or here intentions? Towards whom is he using them? etc.). In this dissertation, we shall focus on the speeches and pragmatics of both economists participating to the standard *versus* behavioral debates and economic agents in

the decision situations (e.g., laboratory experiments, surveys, market transactions etc.) that are discussed in these debates. My contention is that scrutinizing these uses of language can contribute to provide a clearer picture of the main issues underlying the standard *versus* behavioral economics debates.[1]

This general introduction presents three of these main issues that are discussed throughout this dissertation, namely the issue of theoretical unification (0.1), the issue of interdisciplinarity between economics and psychology (0.2) and the positive/normative issue (0.3); before presenting how the intended contributions are articulated in the chapters (0.4).

0.1 Theoretical unification and the three dimensions

In terms of the convenient distinction drawn by Esther-Mirjam Sent (2004) between "new" and "old" behavioral economics, this dissertation is strictly about new behavioral economics. Old behavioral economics stems from the work of Herbert Simon on bounded rationality from the mid-1950s onwards. New behavioral economics (just 'behavioral economics' henceforth) stems from the influence of psychologists Daniel Kahneman and Amos Tversky on Thaler in the late 1970s.[2]

From this date onwards, and mostly by borrowing psychologists' experimental methodology, behavioral economists have investigated empirical regularities of individual behaviors that systematically depart from the predictions of standard models. The diverse behavioral regularities investigated in behavioral economics were once given theoretical accounts by separate models (caricaturing a little, there was one model by regularity), but unifying theoretical explanations have been increasingly sought from around the mid-2000s (one model for several regularities).

[1] For a thorough historical and philosophical perspective on the distinctions presented here and other ones used among linguists and philosophers of language, see Rastier (2012). Mongin (2006b;c) discusses other classical distinctions with respect to economic methodology, e.g., between *sentences* (the linguistic entities), *statements* (roughly, the meanings of sentences in terms of sense and reference), *propositions* (roughly, the meanings of sentences in terms of truth and falsity) and *utterances* (roughly, the actions of pronouncing and writing sentences performed by speakers or writers); or between the *uses* of words and sentences and the mere *mentions* of words and sentences. Other economists have already claimed that language uses play a specific role in economic rationality that is yet to be fully investigated (for surveys, see Zhang and Grenier 2013; and Alcouffe 2013). However, most (if not all) these contributions focus on 'a language' and not 'speech'. The relation between these contributions and this dissertation will not be discussed here but kept for further work, mainly because they are totally unconnected to behavioral economics (though see Chen 2013).

[2] Connections and contrasts between "new" and "old" behavioral economics are not discussed in this dissertation since they have already been discussed at length elsewhere (see, e.g., Sent 2004; Klaes and Sent 2005, esp. p.47; Egidi 2012; Kao and Velupillai 2015; Geiger 2016).

Most models from behavioral economics before the mid-2000s have been reprinted in the volume *Advances in Behavioral Economics* (henceforth *Advances*), edited by Colin Camerer, George Loewenstein and Matthew Rabin in 2004. The structure of one part in *Advances* named "basic topic" gives a nice picture of what behavioral economists take to be the most primary domains of economic analysis. In this part, the contributions are organized into five sections, the second, third and fourth of which are labeled "preferences over risky and uncertain outcomes", "intertemporal choice", "fairness and social preferences", respectively. This way by which three dimensions of economic behaviors – 'uncertainty', 'time', and 'other people' – are saliently *separated* is pervasive in the literature around behavioral economics.[3]

This separation is also present in some classics of standard economics (e.g., Deaton and Muellbauer 1980; Mas-Colell et al. 1995), albeit in a less salient fashion. It is however extremely explicit in the contemporary writings of prominent standard economists on behavioral economics. This is not only the case in the two review-essays of *Advances* by Drew Fudenberg (2006) and Wolfgang Pesendorfer (2006), but also, for instance in the book-length criticism *Is Behavioral Economics Doomed?* by David Levine (2012). Furthermore, consider the position of Fudenberg (2006):

> "behavioral economists (*and economic theorists!*) should devote more effort to synthesizing existing models and developing more general ones, and less effort to modeling yet another particular behavioral observation" (Fudenberg 2006, p.699, my emphasis)

What Fudenberg is asking is what he and Levine (2006; 2011; 2012c) are doing in their recent joint work, by proposing a model that provides a unified explanation of the behavioral economists' regularities in the three dimensions altogether. Hence the theoretical and empirical relations among the three dimensions of economic behaviors is currently an area of economic analysis where the boundaries between behavioral and standard economics are fuzzy, to say the least. One intended contribution of this dissertation is to investigate the conditions under which a theoretical unification (i.e., one model for several regularities) articulating the three dimensions altogether is possible. To do so, we shall scrutinize the uses of language by economists

[3]The two other sections considered as "basic topic" by these behavioral economists are entitled "Reference-dependence and loss-aversion" and "game theory". The former includes contributions about individual behaviors under certainty and will have a specific place in this dissertation. By contrast, the latter includes mostly contributions about strategic interactions (between at least two people) and will not be discussed at length.

in their models of individual behaviors in the three dimensions (e.g., the notion of 'dynamic consistency') and by the economic agents in the situations where the behavioral regularities are observed (e.g., marking the linguistic distinction between 'now' and 'later').

One further clarification is worth stating. In this dissertation, standard economics refers to the economics defended and practiced by self-identified 'standard' economists (such as Fudenberg, Levine, Pesendorfer) that have debated or critically commented on self-identified 'behavioral' economists (such as Thaler, Camerer, Loewenstein, Rabin). However, such criticisms of behavioral economics have somewhat different positions regarding what they take to be the founding principles of standard economics. On the one hand, game theorists such as Levine emphasizes that "[t]he heart of modern "rational" economic theory is the concept of a non-cooperative or "Nash" equilibrium of a game" (Levine 2012, chap.2). Nash equilibrium can be stated as the identification of the rational actions to be undertaken by (at least) two agents independently of one another but the respective consequences of these actions for each agent are mutually dependent on the actions chosen. This dissertation will not be primarily concerned with such game theoretical issues. It will by contrast focus much more on the issues raised by decision theorists such as Pesendorfer for whom the founding principle of economics is the revealed preference "methodology" or "principle" – which are the words they use instead of 'theory' (see Gul and Pesendorfer 2008; see also Dekel and Lipman 2010). For the purpose of this general introduction, revealed preference can be stated as the *identification* of economic agents' choices (directly observable in economic data) with their preferences (indirectly observable in the minds of agents).[4]

For instance, Pesendorfer puts his main methodological concerns with behavioral economics as follows:

> "Traditional choice theoretic models and behavioral [economics'] theories differ in their focus when analyzing "behavioral" phenomena [...]. Theoretical work in behavioral economics often focuses on the psychology of the situation using the economic behavior as a window into the mind of the decision maker [...]. Note the reversal in the roles of economics and psychology. The economic evidence is used to flesh out

[4] A refined view of the embeddedness of 'behavioral', 'standard', 'mainstream' and 'neoclassical' economics is offered by Colander (2000) and Davis (2006, 2007b, 2009b, 2011). On the historical evolution and contemporary state of the relations between decision theory and game theory see, e.g., Mirowski (2002, 2006, 2009) or Aumann and Drèze (2009). On how this relation fits in the evolution of general equilibrium theory, consumer choice theory and demand theory see, e.g., Kirman (1989) or Mirowski and Hands (2006). On the contemporary status of 'revealed preferences' theory, see Hands (2012a; 2013b).

> the details of a particular mental process that is operational for this particular case." (Pesendorfer 2006, pp.718-719)

In other words, the main methodological issue with behavioral economics from a decision theoretic perspective is its relation with psychology; this issue is presented in the next section.

0.2 Economics and psychology: the issue of interdisciplinarity

Pesendorfer is the co-author with Faruk Gul (2008) of one of the strongest and most discussed criticism of behavioral economics in the 2000s, which triggered a whole volume of replies and comments on it: *The Foundations of Positive and Normative Economics* (henceforth *The Foundations*; Caplin and Schotter 2008a). In this volume, the critical exchanges of Gul and Pesendorfer (2008) *versus* Camerer (2008) provide an explicit illustration of the role of economists uses of language in economic rationality with respect to the issue of interdisciplinarity between economics and psychology. At some point in their defense of standard economics, Gul and Pesendorfer put two quotes – one from psychology and one from economics – right next to each others, and then discuss the contrast between them (2008, p.5, references omitted, my emphasis):

> "Much aversion to risks is driven by immediate fear responses, which are largely traceable to a small area of the brain called the amygdala.
> A decision maker is (globally) risk averse, . . . if and only if his von Neumann-Morgenstern utility is concave at the relevant (all) wealth levels.
>
> [. . .]Most researchers recognize the various terms in the second statement as abstractions belonging to the specialized vocabulary of economics. *Though less apparent*, the language of the first statement is equally specialized in its use of discipline-specific abstractions. The terms "immediate fear" and "traceable" are abstractions of psychology and neuroscience."

Notice that Gul and Pesendorfer's point here is not only that economists and psychologists use different languages. It is also that the one from economics (e.g., "von Neumann-Morgenstern utility") is a more *technical language* (or more "discipline-specific") than the one from psychology (e.g., "immediate fear") which is closer to our *everyday language*. In his response, Camerer argues about some behavioral regularities that:

> "many of these phenomena can be translated into conventional economic language. Indeed, a major contribution of neuroeconomics [i.e., which also means 'behavioral

> economics' in this context – DJ] may be to provide a *formal language* for talking about these phenomena. [...] [O]ther phenomena are more clearly understood by adopting new terms from psychology and neuroscience rather than struggling to fit the brain's complex activity awkwardly into the Procrustean bed of economic language." (Camerer 2008, p.50, my emphasis)

In other words, his point is that the *formal language* of mathematics, logic and probability theory used in economics is of interests for both economists and psychologists (and that is "a major contribution" of behavioral economics); but using that formal language should not prevent economists, when they feel the need to do so, to interpret it through the language used by psychologists.[5]

One intended contribution of the present dissertation is to flesh out in a systematic way this perspective on the issue of interdisciplinarity between economics and psychology in terms of the uses of language by economists or psychologists. To do so, I shall use the categories of *formal*, *technical* and *everyday* languages emphasized in the previous paragraphs as follows. *Formal language* refers to the symbolism and syntax of mathematics, logic and probability theory as used for instance by economists to construct the "formal models" Rubinstein was talking about – themselves used for various purposes such as making measurements, drawing inferences, and so on (see Boumans 2005; Morgan 2012). *Technical language* refers to a language used in a "discipline-specific" manner within scientific communities. And *everyday language* refers to our uses of language in our daily lives that have *a priori* nothing to do with the practice of science. I do not mean that these three distinctions cut sharply the whole uses of language by economists and economic agents into neat and independent categories. Quite the contrary, making these distinctions can shed some lights on the standard *versus* behavioral economics debates when we focus on how these categories interact and influence each others. For instance, in economics, the use of formal language encourages to combine the technical term 'convex' from mathematics with the everyday term 'preferences' to make the technical term 'convex preferences' (instead of using a longer everyday version like 'a preference for an average of various goods over an extreme amount of one good'). Historian of psychology Kurt Danziger (1997) provides an account of *the interactions* between the uses of technical and everyday languages

[5]Exactly the same conclusion on the explicit importance of the role of language for standard and behavioral economists could have been made by discussing the two introductions of *The Psychology of Economic Decisions* (Brocas and Carillo 2003b; 2004a).

by psychologists. In his account, he distinguishes '*p*sychology' as the everyday uses of language about the mental (e.g., 'I *prefer* chocolate', 'I had so much *fear*') from '*P*sychology' as the technical uses of language about the mental by psychologists in their scientific activities (e.g., 'This task is meant to put the subjects into a state of *cognitive dissonance*', 'The brain scan shows that the subjects' responses were due to *emotions*', the psychologists' statements in Gul and Pesendorfer's quote above, etc). We shall adopt this typographical distinction between everyday *p*sychology and technical *P*sychology because it is convenient for discussing the issue of interdisciplinarity between economics and *P*sychology underlying the standard *versus* behavioral economics debates. Finally, to better emphasize the interactions between technical and everyday languages, we shall consider them as two instances of *ordinary language* uses, a more general category that I will often contrast as a whole with formal language uses.[6]

Furthermore, the category of ordinary language allows the methodological perspective presented so far to be easily connected to the last issue underlying the standard *versus* behavioral economics debates, namely, the positive/normative issue.

0.3 Positive/normative issues: the entanglement thesis

As the name of the volume mentioned above suggests – *The Foundations of Positive and Normative Economics* – and as the book-length historical perspective on behavioral economics conducted by Floris Heukelom (2014) attests, there are various connections between the issue of interdisciplinarity and the positive/normative issue. Outside of behavioral economics, decision theory and some areas of the methodology and philosophy of economics, the 'positive/normative issue' is rarely associated with models of individual behaviors or rationality. Rather, it is associated with collective forms of behaviors or rationality and the evaluation of market efficiency.

[6]Two remarks are worth making. Firslty, sometimes under different names, these distinctions between different types of language have been discussed by others in economics (see esp. Dennis 1982a; b; 1996; 1998a; b; Vilks 1995; 1998; Weintraub 1998; 2002; Backhouse 1998; Mirowski 2002; 2012; Mongin 2001; 2003; 2004; Armatte 2004; Giocoli 2003), as has been the *p*sychological status of 'preferences' and 'choices' in economics and decision theory (see, e.g., Pettit 1991; Rosenberg 1992, chaps.5-6; Mongin 2011; Guala 2012; Hands 2012a). In both cases, these authors do not tackle the issues underlying behavioral economics that are at the center of this dissertation. Secondly, outside of economics, the issue of interdisciplinarity in contemporary sciences is discussed mainly from in 'interdisciplinarity *studies*' (see, e.g., Klein 2010; Huutoniemi et al. 2010) or regarding the '*unity or disunity* of science' (see, e.g., Oppenheim and Putnam 1958; Kitcher 1984). Cat (2013) argues that both perspectives are sometimes compatible especially in the work of Galison (1999) who focuses on the interactions between languages across and within disciplines; it should be noted that Heukelom (2009) used the latter's perspective on behavioral economics.

At this level, the positive/normative issue concerns the organization of subareas in economics into 'positive economics' and 'normative economics'. Positive economics is supposed to describe, explain and/or predict economic states of affairs, e.g., 'The demand for apples has risen or will rise because the price of apples has decreased or will decrease', such as in consumer choice theory or demand analysis. Normative economics is supposed to evaluate, recommend and/or prescribe economic states of affairs, e.g., 'The production of apples should be subsidized to increase consumers' welfare', such as in welfare economics, public economics or social choice theory. Though models of individual behaviors in economics are rightly seen as 'the toolbox' used by economists to model social phenomena, the positive/normative issue is also present in them through the claims of rationality attached to their interpretations. In standard economics, these models have a positive dimension in the sense that they describe, explain and/or predict *the behaviors of rational individuals*, e.g., 'you prefer apples to oranges and oranges to banana, then *you prefer* or *will prefer* apples to banana, *because* you are rational'. But *the very same models* also have a normative dimension, in the sense they evaluate, recommend and/or prescribe *the rationality of individual behaviors*, e.g., 'if you prefer apples to oranges and oranges to banana then *you ought to* prefer apples to banana, or else you are irrational'.[7]

Can positive economics and normative economics be totally separated? How are they articulated? These are the questions underlying the positive/normative issue at the level of the articulation of positive economics and normative economics. They will be discussed in this dissertation only briefly and only for the purpose of furthering our understanding of the individual level of the positive/normative issue. At this level, the underlying questions are: How is it that the axioms of rationality at the root of these models were considered *both* normative *and* positive and then became positively implausible from experiments but remained considered as normatively sound? Where does the normativity of these axioms and models come

[7]See Hausman and McPherson (2006) or Hands (2012b) on the relations between these two levels of the positive/normative issue. At the individual level at least, it should be noted that the nuance between 'descriptive' and 'positive' is *sometimes* meaningful in economics. The former means something like 'description of human behavioral patterns' and the latter something like 'theoretical explanation of human behaviors'. But that is not always the case. Often *both* 'positive' and 'descriptive' are synonyms and can take *either* of these two meanings. The same ambiguity arises with 'normative' and 'prescriptive', often taken to mean something like 'theoretical account of some norms of human behaviors' (for the former) and 'recommendations for being in conformity with some norms of human behaviors' (for the latter). Also, 'positive' is often more associated with the predictive dimension of a model *as opposed to* the descriptive dimension of its assumptions, following Friedman (1953). In this dissertation, the precise meanings of these terms will be made explicit only when it is crucial for the arguments and when the context does not provide sufficient clues to determine these meanings.

from? In this dissertation, these questions raised in the behavioral *versus* standard economics debates will be scrutinized through the lenses of the entanglement thesis of Hilary Putnam (a philosopher), Vivian Walsh and Amartya Sen (both philosophers and economists). We adopt this thesis mainly because of the central role of both ordinary and formal language uses in its arguments. To understand briefly how, it is worth remembering that the philosophical roots of the positive/normative issue in economics are that "many economists take for granted the positivist tenet that only mathematical and factual judgments – and not value judgments – are amenable to *rational discussion*" (Mongin 2006c, p.258, my emphasis; see also Hands 2001, chaps.2-3; 2012b). Indeed, it seems easily feasible *in-principle* to reach an agreement through rational discussion over what the facts are, e.g., 'the GDP has increased by 10%' or 'he sold one bottle of wine', because the data (and the process of their construction) can checked by others (at least in-principle). The same goes for whether the mathematical proofs of theorems and propositions are corrects, e.g., '$2x = 1 \Leftrightarrow x = \frac{1}{2}$', because the steps of the proofs can be checked by others. But that does not hold for values, especially of an ethical sort, e.g., 'wearing short skirts is morally unacceptable' or 'the riches should be in solidarity with the poor', for which we seem to have no scientific or formal procedures to reach an agreement over. And in any case we feel (as if we shared a meta-value judgment) that we ought to be free to disagree with anybody else's value judgments for a host of different reasons that may please us to entertain. The rebuttal of this positivist tenet along with the proposition of an alternative is the core of the entanglement thesis.

Adapted for behavioral economics, which is needed because this was not Sen, Walsh or Putnam's initial target, the entanglement thesis can be summarized as follows. Whatever scientists and lay people, e.g., economists and economic agents, refer to when they speak about particular facts and particular values cannot usually be fully separated and scrutinized in isolation of one another. But they can be distinguished and their articulation can be scrutinized. Furthermore, their articulation can never be fully understood without taking into account the role of conventions, theoretical ones for the economist and social ones for the economic agents. The uses of ordinary language play a crucial role in the social conventions that are entangled with facts and values in our everyday life as lay people. And somewhat symmetrically, the uses of formal language from logic, mathematics and probability theory play a crucial role in the theoretical

conventions that are entangled with facts and values in science. It is under this version that this thesis will be used to scrutinize the theoretical and empirical contributions constitutive of the standard *versus* behavioral economics debates in order to provide a better understanding of the positive/normative issue within models of individual behaviors.

Roughly, the entanglement thesis is the fruit of a synthesis from various areas of philosophy (i.e., ethics, the philosophy of language, of logic, of mathematics, of mind and of science) that Putnam (2002; 2004; 2013; 2015) worked out before his recent death. Putnam's work (esp. 2002) bears on the positive/normative issues in economics mostly through Walsh's historical and methodological work (synthesized in Walsh 1996). And Sen is taken by both Putnam and Walsh as an illustration of the practice of economic theory self-consciously without the dichotomies, i.e., facts/conventions or facts/values (see esp. Sen 1987; 2002). In turn, Sen acknowledged that Putnam and Walsh's work in the 2000s indeed fleshed out his historical and methodological positions (see Sen 2005; 2009, esp. the fn p.357). In terms of the role of language, three philosophers play a specific role in the background of the entanglement thesis: W.V.O. Quine (in the work of Putnam), Ludwig Wittgenstein (in the works of Putnam and Sen) and John Austin (in the works of Putnam and Walsh). In order to avoid the philosophical dimension of this dissertation overtaking its intended contribution regarding economics, I will not discuss the works of Quine, Wittgenstein and Austin, but stick to ones of Putnam, Walsh and Sen (see also the contributions collected by Putnam and Walsh 2011).[8]

For the same reason, in this dissertation, the entanglement thesis will be used mainly by *illustrating as concretely as possible* various entanglements of facts, values and conventions in the behavioral *versus* standard economics debates. This strategy of working by concrete

[8]The entanglement thesis *brings together* preexisting arguments against the existence of two dichotomies. There are the arguments to the extent that *the fact/convention dichotomy does hold only as a useful distinction*: (1) from the philosophy of sciences by the positivists themselves on the implications of unobservable entities in quantum physics along with the work of Quine from "Two Dogmas of Empiricism" (1951) onwards; and (2) from the philosophy of mathematics and logic about the indispensability of mathematics and logic in science and the consensual agreement over their 'objectivity' despite their absence of reference to observable objects in the sense science is supposed to refer to observable objects from a positivist perspective (see esp. Putnam 1971; 2004). And there are the arguments to the extent that *the fact/value dichotomy does hold only as a useful distinction*: (1) from the philosophical school of pragmatism which argues that regardless of ethical value judgments, the practice of science necessitate a great deal of *epistemic value judgments* such as 'simplicity', 'coherence', 'relevance', 'conservation of past theories' for the construction *and* selection of theories (see esp. Putnam 2002, chap. 2); and (2) from the post-1960s debates in meta-ethics on the distinction between thin predicates that can make value judgments purely evaluative (or 'normative'), e.g., 'good', 'right', 'it is morally good to never lie', and *thick predicates* that can make value judgments that are inseparably both evaluative *and* descriptive (or 'normative' *and* 'positive'), e.g., 'cruel', 'rude', 'rational', 'prefer', 'it is rational to sell a bottle of wine when one does not drink wine'.

illustration seems well in line with the philosophical approach of pragmatism or neopragmatism to which Putnam, Walsh and Sen are often identified and have showed explicit sympathy for. Notably, this strategy seems to fit the two first features used by Wade Hands (2001, pp.216-217) to characterize pragmatism or neopragmatism: (1) the refusal of a foundationalist methodology whereby criteria for the evaluation of scientific contributions are defined *a priori* and from the outside of scientific practice, and (2) the refusal of a sharp dichotomy between theoretical contributions and empirical work in a given discipline. Furthermore, it seems also fit for the intended contributions of this dissertation, as explained in the next section.[9]

0.4 Intended contributions and articulation of the chapters

To better situate the methodological perspective on the role of language in economic rationality presented so far, consider two existing positions on this issue in economic methodology and the philosophy of economics. On the one hand, there is the position of Deirdre McCloskey (1998 [1985]) who analyzed the rethorics underlying economists' uses of language. Roughly, McCloskey is arguing that (1) the norms of conversation in everyday language are necessary and sufficient to deal with (2) the methdological problems faced by economists. Hence (3) the epistemological prescriptions from the philosophy of science sometimes proclaimed by economists, methodologists or philosophers produce parasitic interferences between (1) and (2). On the other hand, there is the position of Don Ross (2005; 2014), which can be seen as the exact opposite of McCloskey's; and by contrast with her, he discusses behavioral economics. Roughly, Ross argues that (1) everyday language produces parasitic interferences in the relation between (2) the methodological issues around behavioral economics and (3) the philosophical solutions he proposes to them from the philosophy of mind and of science (see Ross, Ladyman and Spurrett 2007 for a more general version of his position). The methodological perspective on the role of language in economic rationality developed so far, especially as grounded in the entanglement thesis, seeks a viable middle ground in-between these two extremes. In short, this dissertation takes the position that because (1) everyday language (esp. *p*sychology) is an indispensable

[9]It should be noted that (1) comes also with a refusal of a radical relativism. The differences between pragmatism and neopragmatism are not significant for the purpose of this dissertation, and these should not be confused with the pragmatics of language discussed at the beginning of this general introduction, despite non-trivial existing connections (see Hands 2001, chap.6).

dimension of *both* the economic agent's choices and the economists' theories, paying some attention to it can contribute to *both* (2) and (3). It can contribute to (3) some philosophical problems posed by methodologists or philosophers interested in economic theory or posed by economists themselves in their methodological manifestos. And it can also contribute to (2) some methodological, theoretical and empirical problems faced by economists interested in the individual behaviors of economic agents.

To sum up, the intended contribution of this dissertation is to show how a methodological perspective on the twofold role of language in economic rationality can clarify three main issues (and their connections), underlying the behavioral *versus* standard economics debates: the issue of the theoretical unification regarding the three dimensions of economic rationality, the issue of interdisciplinarity between economics and *P*sychology and the positive/normative issue within models of individual behaviors. Furthermore, this dissertation seeks to go beyond mere clarification regarding the positive/normative issue and the role of language in the behaviors of economic agents. My intention is to provide a constructive criticism of contributions from behavioral as well as standard economists on both of these points. Following the entanglement thesis, it will be argued that both standard and behavioral economists propose an unsatisfying articulation between the positive and normative dimensions of models of individual behaviors; and that recognizing the entanglement of facts, values and convention can actually be theoretically and empirically fruitful. Furthermore, it will be argued that paying some attention to the role of language in the behaviors of economic agents may *sometimes* show that a seemingly irrational behavior can in fact be defended as rational; hence the implicit axiom – known as 'description invariance' – in standard models of individual behaviors preventing the influence of language needs to be weakened (though not dropped entirely), contrary to the positions of most behavioral and standard economists.

This dissertation is structured in five chapters. In the first chapter, the methodological perspective sketched in this introduction is further developed 'in action', that is, the developments are constructed by putting this perspective on concrete problems underlying the behavioral *versus* standard economics debates. The grounding of this perspective in the entanglement thesis is also further developed 'in action', by discussing the criticisms addressed by Sen to the same standard models of individual behaviors that are criticized by behavioral economics. Hence the

chapter is organized as a comparative study of the works of Sen (as a representative of the entanglement thesis) and Thaler (as a representative of behavioral economics). It focuses on consumer choice theory and rational choice theory *under certainty*, where both authors have made their main contributions. Two conclusions of this chapter motivate the subsequent developments in the dissertation. First, despite our focus on economic analysis under certainty, the three dimensions of individual behaviors (under uncertainty, over time and regarding other people) keep popping up when questions related to the causes or determinants of individual preferences arise (i.e., when preferences are not taken as a 'given'). Second, if the role of language in economic agents' choices is to be taken seriously, then the *communicative structure of choices* needs to be made explicit. To do so, two entities are distinguished: the *decision modeler* who poses a decision problem to the *decision maker* who has to make the choice. The decision modeler can, among other possibilities, be an economist, especially in laboratory experiments, where issues related to the interactions between the languages of the economist and the economic agent loom large.

In the second chapter, the challenges posed by behavioral economics to standard economics on the three dimensions of individual behaviors are scrutinized. We structure our discussion around a historical rupture between two periods. On the one hand, the classical challenges posed by behavioral economics from the end of the 1970s to the middle of the 2000s were about the three dimensions taken separately. On the other hand, more recent challenges from the middle of the 2000s onwards are now involving interactions across dimensions. Regarding the communicative structure of choices, it is argued that the marking of linguistic distinctions within the three dimensions (e.g., 'now' and 'later', 'certainly' and 'probably', 'you' and 'me') is a condition of possibility for these challenges to be posed in the first place. Again, two conclusions motivate the subsequent developments in the dissertation. First, historically, there has been a slow shift in standard decision theory from a primacy of the dimension of uncertainty (over the two others) to a primacy of the dimension of time in economists' value judgments of rationality or irrationality about individual behaviors. Second, no models, be they from behavioral economics or standard economics, can explain the full set of behavioral regularities constitutive of the recent challenges from interactions across dimensions. Hence the process of theoretical unification sought by standard and behavioral economists cannot ignore existing

interactions *across* dimensions, instead of focusing on phenomena dimension by dimension.

In the third chapter, a set of models that are explicitly seeking theoretical unification between behavioral and standard economics, which will be called 'dual models', are scrutinized regarding whether or not they can account for interactions across the three dimensions. The role of language in economic rationality is investigated only from the economists' perspective, by looking at the interactions between, on the one hand, the original formal and technical languages from economics, and, on the other hand, a new technical language borrowed from the *P*sychology of self-control. We suggest that the historical shift from a primacy of the dimension of uncertainty to the primacy of the dimension of time observed in the previous chapter finds its paradigmatic illustration in dual models. We also show that only one of these dual models, namely Fudenberg and Levine's, does tackle the issue of the interaction across the three dimensions, hence providing a promising avenue of theoretical unification and reconciliation between standard and behavioral economics. One conclusion motivates the subsequent developments in the dissertation: a set of behavioral regularities known as 'framing effects' is explicitly pointed by Fudenberg and Levine (along with other authors of dual models) as the main limit to their intended theoretical unification.

In the fourth chapter, framing effects are scrutinized to understand their empirical structure and theoretical implications. We focus on one specific type of framing effects, namely those violating an implicit axiom of standard models of individual behaviors known as 'description invariance'. As its name suggests, description invariance is the axiom of standard models that make them blind to the role of language in economic rationality as displayed by economic agents. Hence the role of language in economic rationality is investigated only from the perspective of economic agents, by looking at three decades of research on framing effects in *P*sychology. Because this body of research is not discussed in behavioral and standard economics, the goal of the chapter is to organize it in a way that is useful for an economist who wants to deal with framing effects. It is argued that the received view in economics about framing effects, which is based mostly on the original work from Kahneman and Tversky on the topic in the 1980s, is misleading in several respects. The main conclusion that motivates the last and next chapter is that there is a considerable set of reasons for wanting to weaken the axiom of description invariance, notably because its violations are pervasive in the three dimensions but also because

they are not always irrational.

The last chapter is co-written with a colleague who practice axiomatic decision theory (Dino Borie). We propose an axiomatic framework in which the axiom of description invariance can be made explicit, and then explore how to weaken it in order to account for framing effects. Our interpretation of the mathematical structure in terms of *two* simultaneous but possibly distinct perspectives on *one* decision problem, namely the decision maker's and the decision modeler's, is intended to place the communicative structure of choices in the foreground. We argue that, contrary to the received view on framing effects, the axiom of description invariance is not an 'all or nothing' axiom, i.e., either you have it or you don't. That is, we argue that description invariance can be weakened instead of being totally dropped, and that different weakening imply different degrees of dependence to description. We apply our framework to different framing effects to show either that the few existing models of framing in economics do not deal with violations of description invariance in a satisfactory manner, or that the positive/normative issue underlying framing effect can be clarified through our framework. The simplicity of our framework, along with our results in terms of the classical equivalence between choice, preference and utility, are intended as minor modifications to economists' formal and technical languages in order to take into account economically relevant uses of everyday language by economic agents.

Chapter 1

The entanglement of facts, values and conventions, under certainty

> "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 optimizing 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. For instance, the Pythagorean theorem is a normative theory of how to calculate the length of one side of a right triangle if you know the length of the two sides. If you use any other formula you will be wrong" (Thaler 2015, p.25)

Introduction

In retrospect, 1980 is usually marked as the birth of behavioral economics. The reason is that it corresponds to the publication year of Richard Thaler's "Toward a positive theory of consumer choice" in the *Journal of Economic Behavior and Organization*. In this paper, Thaler

addresses a constructive criticism to standard consumer choice theory. He targets the empirical performance of the theory, which, he argues, makes systematic erroneous predictions in well-identified consumption situations. The constructive aspect of the criticism is that, in order to explain and formalize the correct predictions, he introduces some elements from prospect theory, a theory of decision making conceived by psychologists Daniel Kahneman and Amos Tversky (1979). In other words, Thaler identified systematic deviations between actual and predicted consumption behaviors in economics *and* proposed to account for these deviations by enriching standard economic models with some *P*sychology.

Improving the empirical *and* explanatory scope of economic theory about individual decision making by borrowing insights from *P*sychology are still the main goals of contemporary behavioral economics. Another key aspect of Thaler's work that profoundly marked contemporary behavioral economics is his uses of the words 'descriptive' or 'positive' and 'normative' or 'prescriptive'. He used them to insist that the standard model of consumer choice is descriptive or positive because it supposedly describes, predicts and/or explains consumers' actual behaviors, i.e., what they do, when we assume that they are 'rational'. *While at the same time* the very same model is normative or prescriptive because it *also* recommends, prescribes and/or evaluates consumers' rational behaviors, i.e., what they ought to do to be 'rational'. He further argued that there should be a separation between these two dimensions because psychologists had done so and this is how they got successful at predicting individual behaviors. Though it is now more common since the recent rise of behavioral economics, such uses of these terms were very much unusual at the time *within* consumer choice theory. Rather, they were (and are still) used to distinguish *between* subareas of economics, i.e., positive economics (e.g., consumer choice theory) and normative economics (e.g., welfare economics), *not* to distinguish two dimensions of a single model of *individual* behaviors within positive economics. Indeed, the words 'normative' and 'prescriptive' were (and are still) strongly associated with a *collective* notion of rationality from which, e.g., the efficiency of market allocations is judged.[1]

[1] Here are a few illustrations of Thaler's declarations:

"The economic theory of the consumer is a combination of positive and normative theories. Since it is based on a rational maximizing model it describes how consumers *should* choose, but it is alleged to also describe how they *do* choose." (Thaler 1980, opening sentence of the abstract)

"The problem is in using the same model to *prescribe* rational choice and to *describe* actual choices. If people are not always rational, then we may need two different models for these two distinct tasks" (Thaler 1992, p.4)

This chapter has three related goals, all derived from the perspective presented in the general introduction. The first goal, pertaining to the positive/normative issue, is to make explicit the entanglement of facts, values and theoretical conventions constitutive of some classical challenges posed by Thaler and other behavioral economists to standard models of economic behaviors under certainty. To do so, we shall proceed through a systematic comparison between, on the one hand, these classical challenges from behavioral economics, and, on the other hand, the classical challenges from Sen *to the very same standard models of economic behaviors under certainty*. The second goal, pertaining to the issue of interdisciplinarity, is to study the relations between economics and both *p*sychology and *P*sychology throughout that comparison. The third goal, pertaining to the role of language in economic rationality, is first to show how the challenges from both Thaler and Sen underlie a methodology where the role of ordinary language is non-trivial and then explicitly develop the theoretical and methdological implications of this observation.

More precisely, this chapter focuses on individual decision *under certainty* for three reasons. Firstly, this restriction allows to be more precise on the methodological issues that are at the center of this dissertation so that they will be more easily discussed in the subsequent chapters. Secondly, by contrast with the next chapter which focuses on decisions *explicitly* conceived in economics as being under uncertainty, over time and regarding other people (i.e., not under certainty *per se*), this chapter aims to show *there are already inherent tensions under certainty with respect to these three dimensions of economic behaviors*. Thirdly, this restriction allows a clean comparison of two different critical perspectives – namely Thaler's and Sen's – on the very same models of economic behaviors. These two perspectives are obviously central for this dissertation as Thaler's work is constitutive of the behavioral versus standard economics debates, i.e., the object of this dissertation, and Sen's work is constitutive of the entanglement thesis, i.e., the methodological perspective of this dissertation.

By models of economic behaviors under certainty *per se* I mean the models from consumer

"What should the new kind of economic theory be? The most important advance I would like to see is a clear distinction made between normative and descriptive theories." (ibid, p.197)

"Psychologists distinguish between two kinds of theories: normative and descriptive. To them, normative theories characterize rational choice [...]. Descriptive theories try to characterize actual choices. [...] Economists have traditionally used one theory to serve both the normative and descriptive purposes" (Thaler 2000, p.138)

choice theory *and* rational choice theory. It is worth clarifying what I take to be the relations between these two theories. The principal theoretical motivation of consumer choice theory is to ground the so-called 'Law of Demand' – an inverse relation between prices and quantities demanded of consumption goods – in individual consumers' decisions made under the constraint of their income or budget. From the late 19th century onwards, various authors or even traditions have theorized the consumer's decisions through three main theoretical entities: choice, preference and utility. The meanings and formalizations of these entities have evolved through time and across traditions, the historical and methodological details of which are fairly complex (see Lallement 1984; Walsh 1996; Mongin 2000; and the references in Mirowski and Hands' 2006 symposium). Notably, the theoretical role of money and the budget (or income) constraint has always been problematic regardless of the traditions (see Lallement 1984; and Mirowski and Hands 1998; or Hands and Mirowski 1998). In this chapter, 'rational choice theory' will be used as meaning 'consumer choice theory without a budget constraint and without necessarily involving money'. Hence because the heterogeneity of traditions in consumer choice theory is mainly threefold since the postwar era (Mirowski and Hands 1998), rational choice theory inherits that threefold heterogeneity.[2]

The three main traditions identified by Mirowski and Hands are the general-equilibrium theory developed at the Cowles Commission notably by Kenneth Arrow and Gerard Debreu, the revealed preference theory developed at MIT notably by Paul Samuelson, and the microeconomics from the Chicago School developed notably by Milton Friedman, George Stigler and Gary Becker. A point worth noting for discussing Sen's and Thaler's criticisms is that:

> "Each subprogram had the capacity to absorb certain forms of criticism and thus deflect those criticisms away from the vulnerable areas in other subprograms, where they might do the most damage. Attack general-equilibrium theory for its lack of empirical relevance, and one is quickly directed toward Becker, Stigler, and the grand

[2]There is a wide variety of uses of the expression 'rational choice theory' which do not seem to be all mutually compatible, especially on what they take to be the 'foundational' defining feature of a *rational* choice, e.g., consistency, self-interest, means-ends reasoning, optimization, maximization, preference satisfaction (for discussions of this variety, see Sen 2002, chap.1; Amadae 2003; Herfeld 2013; 2016). By contrast my use of 'rational choice theory' as consumer choice theory without budget constraint is not 'foundational' but merely meant to capture the contrast between, on the one hand, choices in a market with prices and a budget (consumer choice theory), and, on the other hand, choices *possibly* outside of a market without prices and a budget (rational choice theory). We will see below that there is a mathematically non-trivial reason for this distinction (roughly, *all* the axioms underlying rational choice theory in my sense have to be satisfied in consumer choice theory but not all the axioms underlying consumer choice theory have to be satisfied in rational choice theory). I do *not* claim to be the only one using 'rational choice theory' in that sense (see, e.g., Richter 1971; Sen 1971; Bossert and Suzumura 2010, chap. 1; see also Pollak 1990's discussion of "extended domains" of choice).

> Chicago tradition of applied microeconomics. Go after Chicago for its loose use of subjective utility and a priori categories, and one is quickly told how the relevant preferences are operationally defined and could be revealed by direct observations of consumers' choices. Finally, go after revealed preference theory for emptiness as a tautological definition of consistency in choice, and one is immediately shown the way through the equivalence results and right into the heart of the complete Walrasian general-equilibrium model." (Mirowski and Hands 1998, p.289)

In other words, the threefold heterogeneity of traditions confers to standard models of individual behaviors a certain resilience to criticisms. In this chapter we shall keep that in mind in order to fully appreciate the strength of (at least the combination of) Sen's and Thaler's criticisms.

This chapter is organized into four sections, each focusing on a point of comparison between Sen's and Thaler's (together with some other behavioral economists') criticisms. We first highlight how they share the same target, i.e., standard models of individual behaviors under certainty, with the same interpretation of positive/normative issue in those models, but diverge on their methodological positions against that interpretation (1.1). We discuss some methodological statements from both of these two authors displaying a peculiar role of ordinary language for economic rationality, which we propose to flesh out and develop through the so-called theory of speech act (1.2). We then argue that Sen and the behavioral economics of Thaler display a further methodological convergence in their criticism of the exogeneity of preferences in standard economic analysis (1.3). We finally contrast their positions on the articulation of positive economics and normative economics to explain how they rather disagree on *most* points underlying this issue (1.4).[3]

1.1 Same target, same interpretation, but different methodological positions

The goal of this section is to illustrate two counter-examples to standard models of individual behaviors in economics and to explain their theoretical and methodological implications in some details. One counter example from Sen is taken from his work on context-dependent preferences

[3]The chapter could have been organized through a sequential comparison between Sen's and Thaler's (with other behavioral economists') works, i.e., discussing first the contributions from one, then the contribution from the other and finally a comparative synthesis. We have indeed opted for a systematic comparison organized through thematic sections because it fits better the methodological and philosophical (rather than historical) points that are going to be developed. Also we thought that readers acquainted with one of the two authors but not the other would find the systematic presentation more informative than the sequential one.

(1.1.1). The other counter-examples from Thaler is about what he labeled the 'endowment effect' (1.1.2). We then highlight how these two counter-examples come from two symmetrically opposed methodological criticisms of the same interpretation of standard models of individual behaviors (1.1.3).

1.1.1 Sen on context-dependent preferences

Imagine that, at a dinner-guest party, someone does not take an apple from the fruit basket when it is the last one, but he does take it when there is another apple in the basket (Sen 2002, chap.3, p.129). In other words, he chooses *nothing* when he has the choice between $\{nothing, an\,apple\}$ but *an apple* when he has the choice between $\{nothing, an\,apple, another\,apple\}$. Formally, denoting respectively by x, y and z the elements *nothing*, *an apple* and *another apple*, we have a decision maker choosing x from the choice set $\{x, y\}$ but y from $\{x, y, z\}$. Using the formal language of choice functions ($C(.)$), 'x is *chosen from* the choice set $\{x, y\}$' can be written $\{x\} = C(\{x, y\})$ and 'y is *chosen from* the choice set $\{x, y, z\}$' is written $\{y\} = C(\{x, y, z\})$. The conjunction of $\{x\} = C(\{x, y\})$ and $\{y\} = C(\{x, y, z\})$ is an example of what Sen calls context-dependent preferences. Sen uses intuitive examples such as the one with the apple to suggest that context-dependent preferences are perfectly common and rational. How can such a simple choice situation be a counter-example to the grand theories of consumer choice and rational choice in economics?

Despite the simplicity of the example, Sen taps down to the deepest axiomatic grounds of rational choice theory stemming from the revealed preference *theory* of consumer choice (see Bossert and Suzumura 2010, chap.1). In this tradition, axioms are imposed on choices, supposedly observable directly in *factual* data, so as to define a choice function – which, roughly, is a formal summary of the correspondence between possible choices made from possible choice sets. The axioms thus define conditions of internal consistency of choice. The most minimal of such conditions is sometimes called 'contraction consistency': if an element y *is* chosen from a set $\{x, y, z\}$ and belongs to one of its subset $\{x, y\}$ then it (y) *must* be chosen from this subset ($\{x, y\}$). In formal language: $\{y\} = C(\{x, y, z\}) \Rightarrow \{y\} = C(\{x, y\})$. The apple example above violates precisely that condition. Another important and very minimal condition is sometimes called 'expansion consistency': if y *is* chosen from both $\{x, y\}$ and $\{y, z\}$ then it (y) *must* also

be chosen from their union $\{x, y, z\}$. In formal language: $[\{y\} = C(\{x, y\})\&\{y\} = C(\{y, z\})] \Rightarrow \{y\} = C(\{x, y, z\})$. Notice the (emphasized) normativity of formal language here, or, as Sen puts it, "the authoritarianism of some context-independent axioms" (2002, chap.1, p.6).

When these two minimal conditions are imposed on a choice function, it can be proved that the observed choices are made according to a complete and transitive binary relation of preference that is supposedly unobservable and arguably captures how the decision maker subjectively *values* the elements of the choice set. And equivalently, if a decision maker ranks all the objects of choice in the choice set according to such binary relation of preference then the pattern of choices can be generated by a choice function respecting the minimal conditions. And it is well known that a complete and transitive binary relation of preference implies, with an added axiom of continuity, the existence of a real-valued utility function, an important result for the general-equilibrium version of consumer choice theory proved by Debreu in 1954. The "equivalence results" between revealed preference theory and general-equilibrium theory mentioned by Mirowski and Hands (1998, p.289) above can therefore be schematically represented as follows, using $C(.)$ to denote a choice function, $\succsim$ a binary relation of preference ($\succ$) and indifference ($\sim$), $U(.)$ a utility function and $\Leftrightarrow$ the equivalence results:

$$C(.) \Leftrightarrow \succsim \Leftrightarrow U(.)$$

1971 is an important year concerning the first equivalence between choice function and preference relation for both consumer choice theory and rational choice theory. On the one hand, it is the publication year of a volume that fully clarifies various ways of establishing this equivalence for consumer choice theory (Chipman et. al 1971). On the other hand, it is also the publication year of a widely cited paper by Sen (1971) which clarifies the conditions under which the equivalence hold in rational choice theory (i.e., without a budget constraints), which are (roughly) the congruence and expansion conditions exposed above. In other words, those conditions are necessary and sufficient to ensure that: something, say x, is chosen instead of something else, say y, if and only if x is preferred to y and if and only if x gives more utility than y. In formal language: $\{x\} = C(\{x, y\}) \Leftrightarrow x \succ y \Leftrightarrow U(x) > U(y)$. Notice that the formal expression in terms of choice function is the one giving the more information about the *context* of choice, i.e., about the choice set. One implication of Sen's (1971) results is that the internal

consistency of choice conditions for rational choice theory are more minimal than the minimal internal consistency of choice conditions for consumer choice theory; i.e., the former conditions should be respected in the latter (or, put in another way, the latter imply the former), but not *vice versa*. In other words, a consumer choice is always a rational choice but not *vice versa*.

What is worth noting with respect to the entanglement thesis is that these equivalence results have been taken as a scientific warrant to work with *given* preferences without bothering about their content, and still make claims about the rationality of the economic agent but in a value-free way with respect to the agents' own values; thus making the theoretical activity 'value-free'.[4]

In the words of the entanglement thesis: there are some theoretical conventions (the formal structure of the consistency conditions and their implied equivalence with preference and utility) that warrant some value judgments from the theorist (claims of rationality) about some factual observations (the agent's choices). But there is some irony in Sen's contributions to those equivalence results because his long lasting criticisms of revealed preference theory is precisely against that scientific warrant. His intuitive examples, such as the one with the apples above, are meant to show that the weakest internal consistency of choice conditions can be violated in choice behaviors that everybody understand as perfectly rational. Indeed, one can rationally refuse to take the last apple out a fruit basket to respect social conventions of politeness, good manners and whatnot. That does not mean that these conditions cannot characterize rationality in some instances, but that they cannot do so *a priori* and without any reference to something else than choices themselves: "There is no such thing as *purely* internal consistency of choice" (Sen 2002, chap.3, p.127).

The simplicity of Sen's examples can be deceptive regarding the reach of their implications. If economics is supposed to be about rational choices, and the whole formal theoretical systems of rational choice theory and consumer choice theory rest on foundations (i.e., their most minimal

[4]See Walsh (1996, chap. 4; and somewhat relatedly Mongin 2000, p.1143) on how this penetrated in consumer choice theory from rational choice theory. An influential contemporary illustration of the equivalence results being used as a scientific warrant for making value-neutral claims of rationality is to be found in the textbook of Andreu Mas-Colell et. al (1995, chaps. 1-2). Peter Wakker offers some skeptical comments on this scientific warrant (discussing Mas-Collel et al. among others) that is close to Sen's criticism exposed below: "Many authors introducing a new formal principle in a particular theory use naïve and overly broad terms such as rational, coherent, consistent, and so on, for their particular principle to make it appear impressive.[...] Sometimes, such authors go to great lengths to argue that their particular principle is not only necessary, but also sufficient, for rationality. Citation scores show that such naïve terminologies can nevertheless be effective" (Wakker 2010, p.378, fn3, references omitted)

internal consistency of choice conditions) that are plagued with an infinity of counterexamples (rational behaviors violating the minimal conditions), then economic theory is in bad shape. The counterexamples cannot be easily dismissed by formalizing the type of ordinary explanation given above. Indeed, it would be easy to argue that the decision maker is in fact not facing the same objects of choice in the two choice situations: $\{nothing, an\,apple, another\,apple\}$ or $\{x, y, z\}$ in one case or , and $\{nothing, the\,last\,apple\}$ or $\{x, w\}$ in the other. But that strategy brings a host of other problems, roughly because some of the empirical counterpart of the objects of choice is lost: *an apple* and *the last apple*, or x and w, are different formal elements referring to the same thing, whereas in the example above, we had the same formal element referring to the same thing. That loss is due the *interpretation* of an apple *as* the last apple. This is tantamount to the injection of some *p*sychology into the objects of choice, and that is problematic because such *p*sychology is supposed to be confined in the preference relation or the utility function (or more controversially in the choice function), while the objects of choice are supposed to represent the purposeless Nature faced by the decision maker. Notice that the very *individuation* of objects of choice necessitates a bit of interpretation too, because it necessitates at least some perception, e.g., of *an apple* distinct from *another apple* and from the possibility of choosing *nothing*. But if we were to reformulate the choice set as $\{nothing, an\,apple, an\,apple\}$ or $\{x, y, y\}$, then we have the reverse of the previous empirical problem: *an apple* and *an apple*, or y and y, are the same formal element but they are referring to different things (namely to different apples here). Thus the individuation in terms of *an apple* and *another apple* or y and z appears to the least problematic under those considerations, but it happens to be deeply problematic for rational choice theory and consumer choice theory (for more details on the debates around those considerations see the references in Hédoin 2016; Baccelli 2013a; b). Finally, notice also that 'context', usually one of the vaguest notion in social sciences and philosophy, is rather concise in the literature just discussed: it is the structure of the objects of choice in the choice set, i.e., their numbers individuated with minimal interpretation (beyond mere discrimination).

Summing up, Sen targets the deepest assumptions, or '*axioms*', of the standard models of individual behaviors under certainty in economics. He does so by mean of a counter-example that shows, through its intuitive appeal, that the most minimal notion of rationality in economics is not so minimal after all.

1.1.2 Thaler on endowment effects

Imagine someone who bought some bottles of wine at $5 a bottle thirty years ago and refuses to sell even one of them for $100 now (the current market price), though he never spent more than $35 on a bottle (Thaler 1980, example 1; 2015, chap.2). This kind of situation, where the amount of money a decision maker is willing to accept for selling a good x he owns is far more superior to the amount of money he is willing to pay for buying the same good x when he doesn't own it, is an instance of what Thaler calls the endowment effect. Like Sen, Thaler uses intuitive example such as the one with the bottle of one to suggest that endowment effects are perfectly common. Unlike Sen, however, Thaler does not suggest that in addition to being perfectly common they are also perfectly rational, as we shall see. But let's first see how, again, such a simple choice situation is a counter-example to consumer choice theory and rational choice theory.

Thaler's endowment effects are better understood as counter-examples to the rational choice theory stemming from the version of consumer choice theory of "the grand Chicago tradition of applied microeconomics", as Mirowski and Hands (1998, p.289) put it. Unlike Sen's apple story, Thaler's wine story was not designed as a clear-cut logical counter-example to one of the formal axioms constitutive of the structure of consumer (and rational) choice theory. It rather illustrates an anomalous quantitative deviation from a prediction derived from consumer choice theory. The prediction is that if income and substitution effects *are* null for a good, then consumers' willingness to pay for buying the good (if they don't own it) and willingness to accept money for selling the good (if they own it) *ought* to be equal. And the smallest *are* these two effects the closer the two willingness *ought* to be. Notice here that the normativity of formal language comes from the *formal derivation* of this prediction (see Hanemann 1991 and the references therein), compared to the imposition of assumption in Sen's case. Notice also how the consumer's unobservable *values* and his *factually* observable behavior are conflated in his willingness to pay or to accept money.[5]

[5] In other words, Hanemann's (1991) result is that *if* the inverse relation between variations in the price of a good and variations in the consumer's real income (i.e., income effects) and *if* the inverse relation between variations in the price of a good and variations in the relative prices of the good's substitutes (i.e., substitution effects) *are* both null, *then* the willingness to pay or to accept *ought* to be equal. Notice also that these formal results about the role of substitution effect were published (in 1991) after Thaler's 1980 paper, in which substitution effects are therefore not discussed (but are in his post-1991 papers on endowment effects). However, since most of Thaler's examples are about willingness to pay or to accept *the same good*, substitution effects are

In Thaler's example, the decision maker has a (maximum) willingness to pay for wine of $35, on the one hand, but a (minimum) willingness to sell wine over $100, on the other hand, implying a discrepancy of more than $65 between the two. In formal language (with some abuses from the conventional notations in economics), let x denote a bottle of wine, $WTP(.)$ and $WTA(.)$ willingness to pay and willingness to accept functions both isomorphically related to the decision maker's utility function and the symbol $\ll$ meaning 'much more'. Thaler's example can be expressed as $WTP(x) < \$35 \ll WTA(x) > \$100 \Rightarrow U(x) \neq U(x)$, in stark contradiction with the standard model where, quite trivially in this situation, $WTP(x) = WTA(x) \Leftrightarrow U(x) = U(x)$. Notice here that the problem is *not* that we fail to recognize that one bottle of wine thirty years ago and that bottle today are not the same bottle as its value increases with aging. That the bottle was bought thirty years ago is an irrelevant economic aspect of the problem pointed by Thaler. The problem is that today the consumer does not want to pay more than $35 for *any* bottle of wine *and* does not want to be paid $100 or even much more to sell a bottle of wine.

Thaler proposes "that a more parsimonious explanation is available if one distinguishes between the opportunity costs and out-of-pocket costs" (1980, p.44). Hence, abstracting from the explanatory role of the decision maker's income or budget, Thaler moves his target one ground below consumer choice theory, toward Chicago-style rational choice theory:

> "The first lesson of economics is that all costs are (in some sense) opportunity costs. Therefore opportunity costs *should* be treated as equivalent to out-of-pocket costs. How good is this normative advice as a descriptive model?" (ibid)

But let's consider what Thaler thinks of 'how good is this normative advice *tout court*' before. The answer is not as straightforward as it appears at first sight. In their methodological manifestos (e.g., Camerer and Loewenstein 2004), Thaler and other behavioral economists are often arguing that the normative advice from consumer choice theory and rational choice theory under certainty are taken as good ones, though it is never as directly stated as it is for decision under risk, e.g., "[e]xpected utility is the right way to make decisions" (Thaler 2015, p.30). It is rather indirectly stated, as when Thaler describes a list in which endowment effects figures as "Dumb stuff people do" (2015, p.21). In the words of the entanglement thesis, here the theoret-

plausibly null in these examples. This is obviously the case for the wine example, where we are only talking about the decision maker preferences among bottles of wine. Furthermore, Thaler argues that, in this and other examples, income effects are very small, and (blocking a likely objection from the Chicago School) so are transaction costs.

ical conventions ('all costs are opportunity costs') warrant value judgments from the theorists (claims of non-rationality) about factual observations (the agent's willingness to pay and to accept) that contradict the theoretical conventions. That seems to be very different from Sen's critical attitude towards standard economics' characterization of the rationality of individual behaviors. But it will be later argued (in section 1.3 below) that a closer scrutiny of behavioral economists' attitudes (including Thaler's) towards the standard model reveal that they are not always incompatible with Sen's. In any case, Thaler illustrates how 'all costs are opportunity costs' is not a very good descriptive model by many examples where opportunity costs (e.g., of buying extra less-than-$35 bottles of wine from the $100 obtained in selling an old one) are underweighted compared to out-of-pocket costs (e.g., of these extra bottles without selling an old one).

Summing up, Thaler targets the empirical predictions of standard models of behaviors under certainty in economics. He does so by mean of a counter-example that shows, through its intuitive appeal, that much of us have indeed seen numerous such violations of these empirical predictions in our daily life.

1.1.3 Abiding by the fact/value dichotomy *versus* turning it upside down

Sen and Thaler (along with behavioral economics) share the same interpretation of standard models of individual behaviors regarding the positive/normative issue. This interpretation is the one already presented, both in the general introduction of this dissertation and in the introduction of this chapter. That is, on their account, standard models of individual behaviors have both a positive *and* a normative dimension, i.e., they predict, explain and/or describe what decision makers do *and at the same time* they recommend, prescribe and/or evaluate what decision makers ought to do. Thaler's interpretation is quite straightforward and stops at this observation (see the quotes in the first footnote of this chapter). Sen goes a little further. On his accounts standard models are directly normative and indirectly positive. They are directly normative in the sense that they directly characterize the *substance* of rationality, i.e., the constraints for a choice to be rational, such as consistency conditions or self-interest maximization. In other words, they directly states how to "think and act wisely and judiciously, rather than stupidly or impulsively" (Sen 2002, chap.1, p: 42). On the other hand, they are

indirectly positive in the sense that they first characterize the normative substance of rationality "and *then* assum[e] that actual behavior will coincide with rational behavior" (Sen 2007, pp.20-21, my emphasis). According to Sen, in standard economics,

> "[t]he two uses – directly normative and indirectly predictive – are closely linked. Indeed, the latter is basically parasitic on the former, but not vice versa." (Sen 2002, chap. 1, p.43)

Methodological manifestos from behavioral economics (including Thaler's) display a striking disagreement with Sen on how to handle that "parasitic" aspect of the positive *on* the normative. Thaler and behavioral economists tend to interpret the "parasitic" aspect of the positive on the normative as failures of human rationality. They handle those failures by developing positive models of the systematic deviations from standard models that are observed, most of the time, in lab experiments. The positive dimension of standard models is thus taken away from it, only its normative dimension remains, and behavioral economists develop models where the positive dimension is "direct" in Sen's sense. One feature of this methodological move is its underlying interdisciplinary exchanges with *P*sychology, to provide insights for the behavioral assumptions in the modeling of those deviations. That methodological move itself draws on the position of the psychologists from whom the insights are borrowed. Those psychologists, most notably Kahneman and Tversky, come from a tradition where normative standards of behaviors and cognition are used as benchmarks for the *measurement* of various human performances (see Heukelom 2014, chaps. 3-4). This interdisciplinary aspect of behavioral economics' methodology is illustrated in the next section and more fully developed in the next chapter.[6]

By contrast, Sen's position departs from the standard articulation between the positive and normative dimension by, as it were, turning it upside down: he develops the normative

[6] Thaler displays this position and its origin quite vividly in the following passage:

> "The BDR [Behavioral Decision Research] approach to the study of human decision making has been similar to (and strongly influenced by) the psychological approach to the study of perception. Much can be learned about visual processes by studying powerful optical illusions. [Certain deviations from the normative model of individual behavior] have the force of an illusion to many and have provided similar insights into decision processes. It goes without saying that the existence of an optical illusion that causes us to see one of two equal lines as longer than the other should not reduce the value we place on accurate measurement. On the contrary, illusions demonstrate the need for rulers! Similarly, a demonstration that human choices often violate the axioms of rationality does not necessarily imply any criticism of the axioms of rational choice *as* a normative ideal. Rather, the research is simply intended to show that, for descriptive purposes, alternative models are sometimes necessary." (Thaler 1987, p.100)

direct purpose of rational choice theory from its positive indirect purpose. In other words, he theorizes the substance of rationality with insights gained from observations and discussions of actual behaviors 'in the real world' (i.e., not from lab experiments). One underlying feature of this methodological move is its interdisciplinary exchanges with ethics, moral and political philosophy, to provide insights on the plurality of motives underlying individual choices. Sen therefore argues that an explicit account of the value judgments that are at play in actual individual decision making could improve the behavioral assumptions made in positive economics (see esp. Runciman and Sen 1965; Sen 1987; 2002; 2009). This interdisciplinary aspect of his methodology is further illustrated in the rest of this chapter.

Summing up in the words of the entanglement thesis, Sen derives value judgments from observations of factual behaviors to develop theoretical conventions through the scrutiny of social conventions, norms of behaviors and the like. By contrasts, Thaler and behavioral economists use the theoretical conventions of standard economics to warrant claims of non-rationality (as already argued above) *and* to derive factual measures related to individual behaviors.

Conclusion

We have seen that both Sen and Thaler criticize standard models from consumer choice theory with implications for rational choice theory by means of intuitive counter-examples involving situations that are pervasive in our daily lives. While Sen target assumptions and Thaler targets predictions, the main methodological difference is clearly the former's defense of the behaviors in his examples as rational by contrast with the latter's emphasis on their non-rationality. This difference underlies different methodological criticisms of the same interpretation of the positive/normative issue in standard models of individual behaviors. The next sections will keep using the two examples presented here to illustrate some theoretical and methodological points. The next section focuses on the explicit role of *p*sychology and *P*sychology (*p*-&-*P*sychology henceforth), along with the more implicit role of language for economic rationality, in Sen's and Thaler's work.

1.2 *p*-&-*P*sychology in the communicative structure of choices

The goal of this section is to argue that when we look closely at the role of *p*-&-*P*sychology in Sen's and Thaler's (along with other behavioral economists') work, we see that there is what can be called a *communicative structure* underlying the notion of choice, in two related senses. In a methodological sense, there is a communicative structure underlying the relation between the economist and the economic agent: in order to develop his theory, which includes his uses of formal language, the former can speak to the latter through the ordinary language and *p*sychology they share (1.2.1). In an empirical sense, there is a communicative structure underlying the relation between, on the one hand, economic agents using ordinary language to present decision problems to, on the other hand, other economic agents who make choices using ordinary language as well (1.2.2). The implications and limits of such communicative structure for economics are fleshed out by developing some of Sen's methodological reflections through the so-called theory of speech acts (1.2.3).

1.2.1 The methodological communicative structure of choices

A characteristic methodological feature of Thaler's explanations of his examples is that they combine intuitive appeal to the *p*sychology of the reader with some further *P*sychological principle. In the case of the endowment effect seen above, the *P*sychological principle is what is called *loss aversion* in Kahneman and Tversky's (1979) prospect theory: decision makers *value* losses and gains of the same magnitude asymmetrically, losses involving usually twice more disutility than the utility from gains of the same absolute size. That explains the underweighting of opportunity costs in two ways (Thaler 1980, p.44). One explanation is that the disutility from losing *goods* (e.g., removing bottles from one's wine endowment) is superior to the utility from gaining goods (e.g., adding the same bottles to one's wine endowment). And the other non-mutually exclusive explanation is that the disutility of the *money* spent on a good (e.g., \$35 for a good bottle of wine) is superior to the utility of the money received from selling the good (e.g., \$100 for an old bottle of wine) – notice how the consumer's *values* for money is no more taken as observable through behaviors but is subsumed under an unobservable *P*sychological principle. Figure 1.1 provides graphical representations of these two explanations on Kahneman

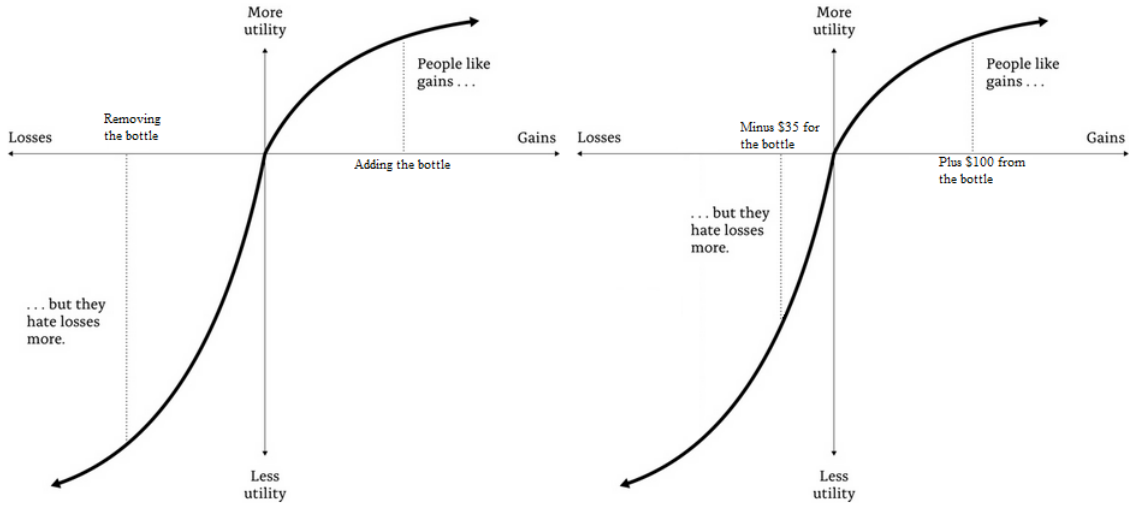

Figure 1.1: Loss aversion and the wine example

and Tversky's 'value function' (through some modifications on Thaler's 2015, figure 3).

The S-shape of the curve represents the asymmetric valuation of, or marginal utility for, losses and gains; and the abrupt change represents loss aversion, i.e., the difference in magnitude between utility from gains and disutility from losses (also represented by the different lengths of the dotted lines). Thaler has expressed the "view of endowment effects and loss aversion as fundamental characteristics of preferences" (Kahneman, Knetsch and Thaler 1990, p.1346).

Loss aversion as a fundamental characteristic of preferences is indeed a distinctive feature of behavioral economics, "the single most powerful tool in the behavioral economist's arsenal" (Thaler 2015, p.33), and a very disputed one from standard economics' perspective. Its conceptual possibility condition, however, the notion of reference-dependent preferences, is (maybe 'paradoxically') not that controversial. Reference-dependent preferences are inherent to prospect theory's most central *P*sychological principle of a *reference point*: a given decision problem can very often (if not always) be presented from (at least) two perspectives whereby a given consequence of a given choice appears either as a gain or as a loss. And when the reference point implies the perspective of losses then by virtue of loss aversion the preference is determined toward the choice that will avoid the losses. For instance, the decision of selling or not selling a bottle of wine can be seen from two different reference points: from (1) a given endowment of bottles of wine, or from (2) the opportunity of buying bottles of wine. But the former presents

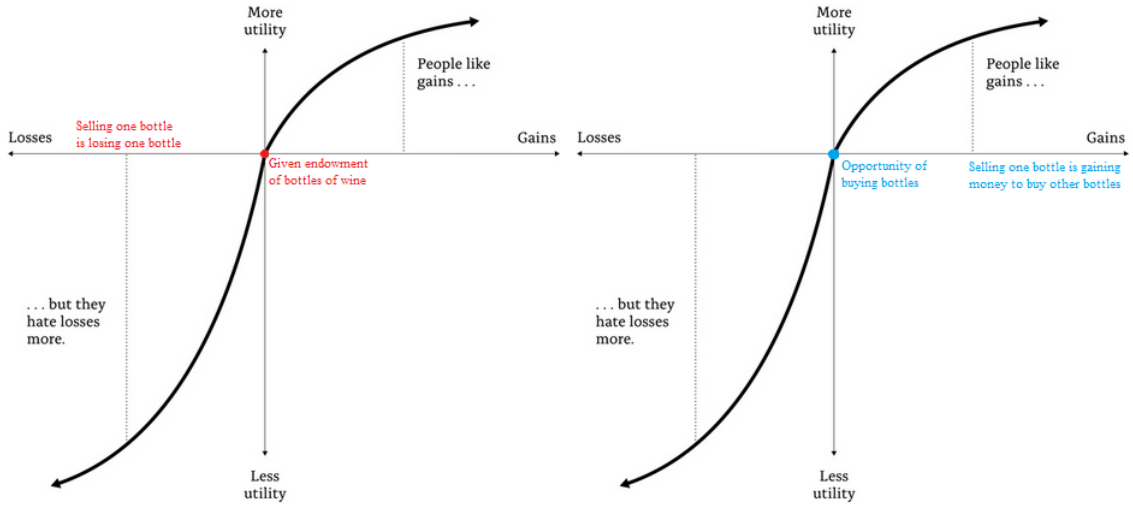

Figure 1.2: Reference-dependence and the wine example

the opportunity of *losing* a bottle of wine as opposed to the latter's monetary *gains* from the sale of that bottle. Hence, in the former presentation, loss aversion determines the choice of not selling the bottle, while the prospect of buying further bottles in the latter presentation determines the choice of selling it. Figure 1.2 provides graphical representations of the reasoning (again through some modifications on Thaler's 2015, figure 3).

Reference-dependent preferences are not to be confused with Sen's context-dependent preferences: a decision problem has always a context constituted by the alternatives available, i.e., by the objects of choice in the choice set. And different contexts – e.g., $\{nothing, an\,apple, another\,apple\}$ as opposed to $\{nothing, an\,apple\}$ – imply different choices – e.g., $an\,apple$ as opposed to $nothing$. Hence the example of the endowment effect discussed so far underlies no context-dependency as there is just one choice set regardless of how the problem is presented, i.e., $\{selling\,the\,bottle, not\,selling\,the\,bottle\}$.

Turning to Sen's *explanations* of context-dependent preferences, his and others' contributions to choice theory consist in making formally explicit the *external* reference to behavior, i.e., reference to the agent's values or to some social norms. Such external references inevitably rely on the shared *p*sychology of the economists and the economic agents, and the connection between Sen's work on rational choice theory with ethics, moral and political philosophy arise at this juncture: his and others' contributions to these areas of philosophy guide and justify his

formalizations of such external references. Obviously, the apple example given previously does not need a grand ethical, moral or political theory to be rationalized. Simple references to the desires to "behave decently [...] without violating rules of good behaviors [...] given [the decision maker's] values and scruples" are enough (Sen 2002, chap.3, p.129). But that is not the case in all of his examples, and the simplicity of such examples have their implications in some preference-based formal systems of ethics, moral and political philosophy, where the minimal conditions of consistency violated in the apple example are imposed *a priori* as universal requirement of rationality, just as in economics.[7]

In their early writings, both Thaler and Sen have defended some methodological theses that acknowledged the role of *p*sychology and communication between the economist and the economic agent. Both subsequently did abide by those theses but without explicitly preaching them as clearly as they did back then. Here are two respective instances where their methodological theses are clearly expressed. Thaler puts it as follows:

> "These examples are intended to *illustrate* the behavior under discussion in a manner that appeals to the reader's intuition and experiences. I have discussed these examples with hundreds of friends, colleagues, and students. Many of the examples have also been used as questionnaires – I can informally report that a large majority of non-economists say they would act in the hypothesized manner [...]. I try to provide as many kinds of evidence as possible for each type of behavior [...] from questionnaires, to regressions using market data, to laboratory experiments, to market institutions that exist apparently to exploit these actions." (Thaler 1980, pp.40-41)

As for Sen:

> "There is, of course, something of a problem in interpreting answers to questions as correct and in taking the stated preference to be the actual preference, and there are well-known limitations of the questionnaire method. But then there are problems, as we have seen, with the interpretation of behaviour as well. [...] The thrust of the revealed preference approach has been to undermine thinking as a method of self-knowledge and talking *as a method* of knowing about others. In this, I think, we have been prone, on the one hand, to overstate the difficulties of introspection and communication, and on the other, to underestimate the problems of studying preferences revealed by observed behaviour." (Sen 1973, pp.257-8, my emphasis)

[7] On the whole axiomatic machinery of external references see Sen (2002, part II), Bossert and Suzumura (2010, chap.8) and the references therein. On the implications of simple examples such as the apple one to preference-based formal systems of ethics, moral and political philosophy see Sen's (1983, section VII) discussion, along with most of the discussions in parts III and IV of his 2002 volume.

Thaler the economist talking with his "friends, colleagues, and students" is just illustrating Sen's point on "talking *as a method* of knowing about others" (my emphasis). To be sure, at a certain level, that is unavoidable for any theorist (especially working on individual decision making) in standard or behavioral economics. My point is just that the communicative structure between the economist and the decision maker is explicitly acknowledged by Thaler and Sen in a way that goes further than the often recognized role of so-called 'casual empiricism' in economic theory.

Summing up, the uses of *p*sychology and/or *P*sychology by Sen and Thaler make their explanations of counter-examples to standard models of individual behaviors appealing notably because of a methodological communicative structure underlying the relation between the economist and the economic agent. In the next section, we shall see that 'in the real world', so to speak, this is paralleled by an empirical communicative structure underlying the making of economically relevant choices.

1.2.2 The empirical communicative structure of choices

The goal of this subsection is to argue that, beyond the realm of theory making, the communicative structure illustrated in the previous subsection has an empirical counterpart in the realm of decision making. We shall illustrate this claim through four sets of related issues that are relevant to contemporary debates between standard and behavioral economics.

The first set of issues arise with force in contingent valuations studies. The latter consist in elicitations of individuals' willingness to pay or willingness to accept (mostly) for non-market goods. Discrepancies between willingness to pay and willingness to accept is a well-known problem in contingent valuations – the latter being far greater than the former, just as in endowment effects. One underlying problem that is often silenced in discussions of those discrepancies within the contingent valuations literature is that it is not rare to observe more than fifty percent of protest from the decision makers who are being under elicitation of willingness to accept (Kahneman, Knetsch and Thaler 1991, pp.202-203). Protesters manifest themselves by refusing to state their willingness to accept or by stating unusual answers (refusal to sell or asking literally an infinite amount of money).

> "These extreme responses reflect the feelings of outrage often seen when communities are faced with the prospect of *accepting a new risk* such as a nuclear power plant or waste disposal facility. [...] [C]ompensation [...] often do not help as they are typically perceived as bribes. This is a situation in which people loudly say one thing and the theory asserts another. It is of interest that the practitioners of contingent valuation *elected to listen to the theory, rather than to the respondents*." (Kahneman, Knetsch and Thaler 1991, pp.202-203 and fn4, my emphases)

This is a vivid illustration of a conflict between the influence of the economist's formal language on his ordinary language, or more precisely on the "practitioners"' ordinary language, which blinds them to the agents' ordinary language, i.e., "the respondents" who were not being listened to. In other words, the formal language from economic theory creates some noises in the exchange of information that takes place within the communicative structure underlying the relation between an economic agent who poses a decision problem to another economic agent, i.e., a representative of an institution asking a citizen to state a willingness to pay or to accept. There are obvious concrete economic implications here. Although Thaler and his co-authors feel concerned about this problem, they do not propose a clear solution to it. What they do instead is very close to the work of Sen: they go on discussing the implications of loss aversion concerning *social preference* for fairness. That is a topic discussed in the next chapter, but notice here how an anomaly regarding *risk* (emphasized in the quote) suggest explanations from another dimension of economic behavior, namely social preferences and norms.[8]

The second set of issues pertain to experiments conducted in laboratories. Thaler's endowment effect, especially through the experiments he conducted with Kahneman and Jack Knestch (Kahneman et al. 1990), is part of an ongoing controversy between standard and behavioral economics. Roughly, the controversy is about whether we really need the language of *P*sychology in explaining the phenomenon or whether we can either find a way out with the one from standard consumer choice theory or cancel the effects through more traditional experimental designs from economics (showing the phenomenon was just an artifact after all). It has been suggested that subjects may try to infer information about the quality of the good they are endowed with from

[8] See Sandel (2000) for a critical discussion of several cases other than environmental matters where bringing money into non-market goods triggers the same reaction. See Milanesi (2010) and the references therein a historical perspective on the method of contingent valuations in economics and further ethical issues involved in its application to the environment in economics. Beyond the environment, see Weber (1996) for a balanced criticism of how the Beckerian theory of household consumption blinds practitioners conducting surveys on domestic consumption to the 'real' subjective cost-benefit analysis and opportunity costs perceived by economic agents (the role of *time* is central in this misperception). I thank Maxime Desmarais-Tremblay, Tom Juille and Agnès Gramain, respectively, for having pointed these three references to me.

the mere acts of *the experimenter choosing* to endow subjects with the good. If that is the case then the latter's social preferences may play a role if the endowment is seen as a gift given by the experimenter (see Wilkinson and Klaes 2012, p.212). In other words, the formal language of economic theory may not be translated faithfully enough in the experimental instructions, leading to experimenter effects, i.e., different interpretations of the same instruction from the experimenter and the subjects (see Bardsley 2005; Zizzo 2010 for precise accounts of experimenter effects). Hence endowment effects may be due to some noises created in the exchange of information that takes place within the communicative structure underlying the relation between, on the one hand, an experimenter posing a decision problem to, on the other hand, the subjects.[9]

That such experimenter effects are obviously generalizable to all behavioral experiments has been part of yet another controversy around behavioral economics in the 2000s stemming from the criticisms of Chicago School microeconometricians Steven Levitt and John List (2007a,b; 2008). In his reply to them, Camerer (2015, esp. pp.260-263) argues that these are well-known problems in the behavioral sciences with which behavioral economists have dealt from the start, notably by using "very carefully written" (p.260) experimental instructions. It is therefore not an issue to be confused with the social preferences of the decision makers, a theoretical construct meant to capture something else. Without disagreeing with Camerer's position, I would like to highlight how experimenter effects *and ways of correcting them* remain a methodological issue where the role of language is deeply entrenched. This is so for a quite simple reason: there is *necessarily* some sort of communication between the scientist and its subjects in the lab because experimental instructions and procedures are *necessarily* written or spoken. The

[9]If such a controversy is raging, that might be because the theoretical and practical stakes concerning elicitations of consumers' willingness to pay and to accept are quite substantial (as in contingent valuation studies, where experimental results can be used to correct some elicitation problems; see Robinson and Hammit 2011a, b). On the controversy over endowment effects see Wilkinson and Klaes (2012, Case 5.1), and Isoni et al. (2011) *versus* Plott and Zeiler (2011) for the latest round; see also Engelman and Hollard (2010) for an interpretation of, and experimental evidence on, the endowment effects as driven by various forms of uncertainty subsuming social preferences. By "more traditional experimental designs from economics" the experimental methods developed by economists in the study of market efficiency by contrast with the ones developed by behavioral economists which are inspired from *P*sychology; roughly, the former but not the latter necessarily respect the following three conditions: systematic use of monetary incentives, anonymity thoroughly ensured, subjects not being deceived by experimenters. With Nicolas Vallois (Jullien and Vallois 2014), we have argued for a threefold clarification between 'behavioral economics', 'experimental market economics' (the two approaches just mentioned) and 'experimental economics' (an hybrid of the two former). Keep in mind that those are methodological distinctions *between approaches, not authors*; although some authors have contributed more to (or even constituted) some approaches, most authors have contributed to all approaches. See Cot and Ferey (2016) and the references therein for historical and methodological discussions of experimental economics.

following anecdotal evidence illustrates the relevance of this point for the present discussion. Experimenters in economics always run a "pilot" of their experiments with colleagues instead of 'real' (i.e., paid) subjects in order to check, among other things (e.g., if the computer program runs well), if the experimental instructions are clear enough (see Guala 2005, chap.2). In one such pilot to which I participated, the experimenter had written the symbols '$\{\emptyset\}$' in one decision problem. A non-economist colleague not acquainted with mathematics and set-theory did not get (I quote from memory) 'what those symbols meant' and therefore did not understand that some choices implied *nothing* as consequences. Discussing this issue during the debriefing, the experimenter was surprised *and* quite disappointed that he had to change the symbols for something more comprehensible because (I quote from memory) 'the symbols did match exactly with the model'. Again, this is a case where the formal language from the theory creates some noises in the informational exchange within the communicative structure of choices.

The third set of issues is about economic choices that do not necessarily involve an economist (at least directly) in the communicative structure of choices, e.g., in a market. An instance of this set of issues is raised in one of Thaler's (1980) example of a market institution (in the U.S.) trying to exploit the endowment effect. The example concerns a bill that was about to pass in Congress to outlaw existing agreements between credit card companies and stores whereby the former prevented the latter to charge higher prices for credit card payments. During the negotiation of the bill, "the credit card lobby turned its attention to form rather than substance" by agreeing on such a bill if the stores presented the difference between cash and credit card payments as cash discounts rather than credit card surcharge (1980, p.45). The point is that when the stores actually charge higher for credit card payments, the underlying economic consequences for the consumers are the same whether it says 'cash discount' or 'credit card surcharge'. In both cases it is cheaper to pay by cash and more expansive to pay by credit card. The only difference is in how the *presentation* of the prices is *written*. But the 'surcharge' presentation clearly discourages the uses of credit card, implying rather unwelcome economic consequences for the credit card companies, who therefore lobbied for the consumer's economic decisions to be presented as 'cash discount'. Whether or not this creates noises in the informational exchanges between consumers, stores and credit card companies is arguably a matter of which of these three agents' perspective one takes.

In short, consumers confront their decision problems as they are presented in a certain way through certain words that are not arbitrarily chosen. Some of these words are even constitutive of the 'objects of choice' faced by the decision makers, i.e., the empirical counterpart to the theoretical entities represented in a choice set. For consumer choices, this issue is obviously related to the careful packaging that constitutes the presentation of much consumption goods, just as words are constitutive of the options among which they have to choose in questionnaires or lab experiments, or the answers they give as statements of their choices. Understanding or interpreting such words is therefore constitutive of the process by which most (especially economically important) decisions are taken; and subsequently using such words is directly or indirectly, by way of speaking or by way of writing, constitutive of most (again, especially economically important) *acts of choice*:

> "Every economic choice (even institutional choices) depends on an individual saying "Yes," nodding, handing something to a cashier, signing a contract, reaching into a wallet, clicking "submit" online, releasing an earnings announcement, or executing some other action that requires brain activity." (Camerer 2008, p.47)

I think it is uncontroversial that all these are actions embedded in a communicative structure, *because* it is uncontroversial that economic decisions have informational structures (which receive a substantial share of mathematical economics' attention) with senders and receivers of messages. Taking stock of that communicative structure could not only refine informational analyses but *also* behavioral economists' analyses of decision making since the psychological processes underlying human language is one of the most studied topics in the cognitive sciences.

The last set of issues can be thought of as pushing the previous one a bit further to concentrate on information *per se*. The media (press, TV etc.) seems a paradigmatic example and the underlying issues can be illustrated by a set of field and experimental studies conducted by psychologists Marwan Sinaceur, Chip Heath and Steve Cole (2005), and used by Camerer (2008) in an instance of the behavioral *versus* standard economics debates. It concerns the impact on consumers of a change in the labels used by the media in France during the Mad Cow Crisis, from "bovine spongiform encephalopathy" to "mad cow disease", both referring to the same disease. Market data showed a decline in beef consumption and experimental results showed more emotional reactions to the "mad cow disease" label than to the other one. Camerer's point

is that those phenomena are easily describable "in the language of preferences, beliefs, and constraints" by inferring "from the data, for example, that [...] that relabeling BSE as "mad cow disease" genuinely conveys new information (*doubtful*) or grabs attention and activates emotion (*more likely*)" (Camerer 2008, p.50, my emphases).

But it can be argued that it is not because there is an attentional and emotional dimension in the agents' attitude that the communicative structure underlying the conveying of information is irrelevant in explaining the former dimension. For the emotional responses come from the informational structure and its underlying communicative structure – information doesn't, as it were, fall from the sky. It can be furthermore argued that the very fact that the medias *chose* to change the words *is* a genuine new information that grabs attention and activates emotion differently depending on one's attitudes toward the media outlets (see Sen 2009, p. 336 on the informational and value-formation roles of the media).

Summing up, the methodological communicative structure highlighted in the previous subsection seem to have rather pervasive empirical counterparts, from the uses of questionnaires, to lab experiments, market institutions and the media. In each cases, the communicative structure is constituted at least by an agent who poses a decision problem in ordinary language to another agent who has to make an economic *choice*. The sharing of a common *p*sychology is a condition of possibility for such communication to be possible at all, let alone to be efficient. As Nicholas Bardsley (2008, p.131) puts it about experimenter effects, they "ha[ve] methodological implications but also potential theoretical consequences, since standard accounts of such effects invoke a more sociological model of action than economists have traditionally employed". The next subsection sketches a way to take the role of ordinary language uses into such a model of action without going too far away from economics.[10]

[10]Besides the four sets of issues discussed in this subsection, other ones could have been the verbal and non-verbal means of communication in financial markets as discussed by Schinkus (2010) or the activity of communication by central banks in their conduct of monetary policy as discussed by Blinder et al. (2008). Both are however less directly connected to consumer choice theory and rational choice theory than what was discussed here

1.2.3 The communicative structure of choices and the theory of speech acts

The goal of this subsection is to flesh out the theoretical consequences of the communicative structure of economic choices in a way that balances logical and sociological considerations.

The discussion of the examples from behavioral economics in the previous subsection was in line with Sen's position that "[a]n act of choice [...] is, in a fundamental sense, always a social act" (1973, p.252). And although we are not constantly being "aware of the immense problems of interdependence that characterize a society", behaviors are not "mere translation of [...] personal preferences" (ibid). One of the underlying methodological implications for rational choice theory of acts of choice always being social acts is that choice should not be thought to have the same logical structure than statements or propositions. As Sen puts it:

> "Statements A and *not-A* are contradictory in a way that choosing x from $\{x, y\}$ and y from $\{x, y, z\}$ cannot be. If the latter pair of choices were to entail respectively the statements (1) x is a better alternative than y, and (2) y is a better alternative than x, then there would indeed be a contradiction here." (Sen 2002, chap.3, p.126)

But:

> "An act of choice is not a statement of any kind. Nor does it entail a statement, on its own." (Sen 1995, p.27)

And the Senian motto of external references resurfaces: it is only through them that we can interpret choices as implied statements subjected to the logical rules of contradiction and entailment that are constitutive of the formal language of logic. Sen further illustrates his point by making a very brief parallel with ordinary language uses:

> "The statements A and *not-A* do make a contradictory pair; the acts of saying them need not. Indeed, being consistent or not consistent is not the kind of thing that can happen to choice functions without interpretation – without a presumption about the context that takes us beyond the choices themselves." (Sen 2002, chap.3, p.127)

In other words, although acts of choice do not share the same logical structure as statements, they do seem to share the same logical structure as the acts of saying those statements, at least with respect to contradictory relations. This point can be fleshed out through the so-called

theory of 'speech acts', one of the main concern of which is to work out the similarities and differences between statements and acts of saying statements.[11]

We shall here follow John Searle's (1969; 1979) version of the theory of speech acts. The starting point of the theory is that the act of saying something is called a 'speech act' and is taken as the unit of communication. At the most general level, speech acts are formalized by $F(p)$. Following Searle's (1979, chap.1) terminology, F represents the "psychological state" expressed while speaking, and p the proposition that represents what is said (to be loosely associated with 'statement' in Sen's terms). Using this formula, we can easily illustrate Sen's point about the difference in logical structure between propositions and the acts of saying them. Imagine that a representative from a credit card company in Thaler's example wants to say that the company prefers that their customers use their credit cards. At a basic level abstracting from contextual features, the following four speech acts $F(p)$ are perfectly consistent with each other:

"F_1 = We prefer *that* (p_1 = our customers pay more in credit card)"

"F_2 = We **do not** prefer *that* (p_2 = our customers **do not** pay more in credit card)"

"F_2 = We **do not** prefer *that* (p_3 = our customers pay less in cash)"

"F_1 = We prefer *that* (p_4 = our customers **do not** pay less in cash)"

Notice the contradictions between p_1 and p_2, and between p_3 and p_4. Yet all four speech acts $F_i(p_i)$ are consistent ($i = 1, 2, 3, 4$). That can be noticed only if one pays attention to something *external* to the proposition, namely the psychological states F of the speaker, which is here explicit in the act of saying though it needs not be (especially in daily conversations). Arguably, the main advantage of the $F(p)$ formula is that it can capture quite simply the two main aspects of the communicative structure discussed in this section, namely the shared *p*sychology (F) and the use of ordinary language (through p).

[11] On the place of speech acts within philosophy of language and the sciences of language more generally see Smith (1990), Recanati (2006), Bach (2006), Sadock (2006), and Green (2014). One specificity of that theory, at least as conceived by two of its instigators – Austin (1975 [1955]) and Searle (1969) – is that it considers the philosophy of language to be a subarea of the philosophy of action (which seems like an appropriate starting point to integrate some philosophy of language for rational *choice* theory), itself a subarea of the philosophy of mind for Searle (which seems even more appropriate regarding the rise of behavioral economics). Searle and Vanderveken (1985) provide an axiomatic treatment of Searle's theory of speech acts which uses (most of) the same formal language used in economics. Speech act theory has been discussed in relation to economics regarding the so-called 'performativity' of some economic statements by economists or economic institutions, i.e., when the description of economic states of affairs creates a part of these very states of affairs and further ones (see, e.g., Mäki 2013 and the references therein). These issues are beyond the scope of the present dissertation.

Two limits of the scope of this formula should be acknowledged. The first limit is that the more contextual features one adds, the greater the number of speech acts that can be made consistent with each other. Indeed, besides the speaker's psychological states, the situation in which the speech act is to be performed greatly influences the proposition that the speaker will *choose* to express. For instance, in the above example, it is very likely that the representative performs the first speech act if he is addressing some shareholders, but not if he is addressing a consumer union, in which case he'll probably go more for something confusing like the second or third ones – or yet add various contextual features (some suggestive details are offered by Kitch 1990 on the whole story behind Thaler's credit card example). Beyond the words and the propositions they express, what matters more is *that* they are expressed, *how* they are expressed and *who* are expressing them *to whom*. Formal languages are notoriously at pain to capture those features, which are nevertheless constitutive of the daily business of ordinary language. Hence the formula $F(p)$ is not saved from such difficulties, though most of them are explicitly acknowledged and addressed by linguists (pragmaticians) and philosophers of language working on speech act theory.

The second limit is that the communicative structure of choices depicted so far includes at least two speakers: the one who presents the decision problem *and* the decision maker who makes the choice. Economics (including Sen, Thaler, behavioral economics, rational choice theory and consumer choice theory) is primarily concerned with the acts of choice of the decision maker. Sen suggests that there is a logical similarity between the act of saying and the act of choosing. Indeed, notice how the four examples above can easily be interpreted as four ways by which a decision maker expresses his preferences with contradictory propositions though the overall behavior, i.e., the economic choices or set of speech acts, is logically consistent. However, it seems that to replace the study of the decision maker's preferences by the study of his speech acts in interactions with the speech acts of the one who posed the decision problem is too radical. Tractability issues and/or ending up not doing economics anymore are looming large.

While the first issue is arguably inherent to any theory of action, I would like to sketch a way of addressing the second one. The key distinction to draw, I suggest, is between the decision maker who makes the choices and what will be called in the rest of this dissertation the *decision modeler* who presents the decision problem to the former. The decision modeler *can sometimes*

be identified with an economist or a decision theorist, as in lab experiments or questionnaire surveys. But that is not necessarily the case, and most of the time the decision modeler is to be identified, in the real world, by various social entities (firms, banks, administrations, the media etc.), corresponding to what Thaler recently came to label "choice architects" (discussed in section 1.4 below). That necessity of a decision modeler always involved in a decision problem goes against the background assumption in decision theory of a decision maker facing the brute force of a purposeless 'Nature', and it seems that the framework of game theory would be much better suited to discuss those issues. That seems like a reasonable way to go when the interactions between the decision modeler and the decision maker, i.e., the communicative structure, is strategic. But notice that it is *not necessarily* the case in the examples presented in the previous subsection. The axiomatic framework for decision theory constructed (with Dino Borie) in the last chapter of this dissertation will be primarily motivated, in terms of interpretation, by the distinction between a decision modeler and a decision maker. Until then, that distinction will be discussed only informally. The idea is that decision makers' preferences, choice and utility can be discussed as usual in economics, but keeping in mind that his preferences in a given problem are most of the time conditional on a given speech act from a decision modeler (this could be formally represented as '$\succsim |F(p)$'). This captures the following remark made by Sen (1995, p.25):

> "Observation is not a one-way process. Just as the decision theorist "reads" what people choose, people also "read" what is being offered."

There is a methodological reason for focusing an account of the communicative structure of choices on the decision modeler: the kind of control we can hope to have concerning the observability of the data needed for potential modeling within this perspective. Notably, we have a nearly total control over the speech acts performed by the decision modelers in experimental lab since he *is* the economist and that most of the instructions that are presented to the decision makers are neatly recorded (usually for the purpose of experimental replication, see Camerer 2015, pp.256-257). By contrast, although we can have a fairly high degree of control over the decision maker's speech acts (especially in the lab), it should be acknowledged that not all economic choices in the classical sense of the term of, e.g., buying something, *necessarily*

involves the performance of a speech act strictly speaking. Recall, for instance, Camerer's daily examples of "nodding, handing something to a cashier, [...] reaching into a wallet" (2008, p.47) where the decision maker doesn't speak even though he is involved in a transaction and thus within a communicative structure.[12]

Summing up, we have proposed to characterize some theoretical implications of the communicative structure of choice presented in the previous subsection through the theory of speech act. For these implications to be related to economics, we have, on the one hand, followed Sen's reflections on the distinction between acts of choice and statements about acts of choice, and, on the other, characterized the communicative structure of choices with a 'decision modeler' whose role is very close to Thaler's introduction of 'choice architects' in behavioral economics.

Conclusion

This section developed two related senses of the communicative structure of choices, a methodological one between the economist making theory and decision makers making choices and an empirical one between a decision modeler posing decision problems and a decision maker making choices in those problems. We have suggested that both senses are in line with Sen's and Thaler's (with other behavioral economists') works, which we have used to develop some implications of the communicative structure of choices through the theory of speech acts. Furthermore, it was emphasized that sharing both a common *p*sychology and a common ordinary language is a condition of possibility for the communicative structure of economic choices. The role of *P*sychology in the communicative structure of choices was not discussed very much. This will be addressed in the penultimate chapter on framing effects. Until then, the communicative structure of choices will be developed throughout this dissertation as it will be used to shed some new lights on the classical issues raised by behavioral economics. The next section discusses one such issue about the notion of 'given' preferences in standard economics.

[12]However, these are still intentional actions. The logical structure of speech acts discussed above is derived from the logical structure of Searle's theory of intentionality, the formalization of which is accordingly very similar: $S(p)$ represents whole intentional states with S standing for various psychological modes. Hence the same arguments presented here about Sen's distinction between the (*decision maker's*) acts of choices and statements about such acts of choices carries over non-linguistic acts of choice. A detailed account of the decision maker's intentionality within that framework is however beyond the scope of this dissertation (I have presented such an account elsewhere, see Jullien 2013).

1.3 Convergence against 'given' preferences

The goal of this section is to compare the criticisms addressed by Sen and the behavioral economics of Thaler to one of the deepest methodological convention or "methodological conviction" (Mongin and d'Aspremont 1998, p.384) in standard economic analysis, namely that preferences are *given*, i.e., they are fixed or exogenous variables in the explanation of economic situations under analysis. I will suggest that Sen and behavioral economists have produced criticisms of given preferences that are not mutually exclusive. These criticisms, although initially about decisions under certainty (1.3.1), drift towards either one of three dimensions of economic behaviors beyond certainty and take this dimension to be a *primary* determinant of individual preferences. It can be argued that we naturally drift towards a primacy of *time* preferences in the work of Kahneman without Tversky (1.3.2) and towards a primacy of *risk* preferences in the work of Tversky without Kahneman (1.3.3) – Thaler's and other behavioral economists' work are directly or indirectly related to both drifts. Finally, we naturally drift towards a primacy of *social* preferences in the work of Sen (1.3.4). Though the labels 'primacy of social or time or risk preferences' are somewhat inadequate regarding the issues discussed in this chapter (for reasons explained below) we shall keep using them because they provide a nice point of connection with the next chapter (which focuses exclusively on these three dimensions).

1.3.1 Violations of invariance: procedure, description and context

Claims of behavioral economists against the methodological convention of given preferences are usually made with respect to phenomena that violate three axioms of standard models of individual behaviors: "procedure invariance", (2) "description invariance" and (3) "context invariance" (Camerer 1995, sections H-I and fn.55). Violations of these three axioms show how preferences may not be inherent to decision makers but dependent on, respectively:

(1) how they are elicited, e.g., willingness to pay for a bottle of wine *versus* willingness to accept money for (selling) a bottle of wine;

(2) how parts of the objects of choice are described, e.g., 'cash discount' *versus* 'credit card surcharge';

(3) how the choice set is structured, e.g., $\{nothing, an\,apple, another\,apple\}$ *versus* $\{nothing, an\,apple\}$;

From a historical perspective, (1) and (2) were not explicit axioms of standard models, unlike (3). Indeed, the latter corresponds to the consistency conditions imposed on choice functions in the revealed preference tradition, e.g., expansion and contraction consistency – the targets of Sen's criticisms presented above. Starting around the mid-1980s (esp. since Tversky and Kahneman 1986), observed violations of (1) and (2) had two main consequences. A theoretical consequence was to point out the existence of such implicit axioms in standard models. A methodological consequence was to ground Thaler's (1980) early argument for the separation of the normative and positive dimensions of models of individual behaviors. Indeed, behavioral economists share the view of the psychologists who first demonstrated violations of (1) and (2), namely that these two axioms are "normatively unassailable and descriptively incorrect, [hence] it does not seem possible to construct a theory of choice that is both normatively acceptable and descriptively adequate" (Tversky, Slovic and Kahneman 1990, p.215). It should be noted that in Tversky's landmark contributions to (3) with Itamar Simonson, some violations of it are discussed as rational when relevant information can be rationally inferred from the very structure of the choice set, quite like in Sen's contributions to (3). But some other violations are also discussed as not being normatively defensible. Again, this is put as an argument for the claim that no model of economic behaviors can be *both* positive and normative *at the same time* (see Simonson and Tversky 1992, p.284; Tversky and Simonson 1993, pp.1181-2; Shafir, Simonson and Tversky 1993, pp.25-26; Tversky 1996, p.194).

Behavioral economics has triggered a fair amount of contributions on (1) and (3) in the 2000s. These contributions are done by behavioral economists themselves *but also by standard economists* (see esp. Giraud 2012 regarding (1); Ok et al. 2015 regarding (3); and the references therein). In the latter, quite like in Sen's contributions to (3), preferences are not totally given in so far as they depend on features that are external to the decision makers' preferences about the material consequences of choices. Unlike Sen's contributions, however, these axiomatic characterizations are either agnostic about the rationality or irrationality of the departures from standard models they are formalizing (e.g., Giraud 2012) or emphasize that they are formalizing departures from rationality (e.g., Ok et al. 2015). In that latter case, the statistical normality

of certain behavioral patterns (i.e., the 'facts') don't shake up the normativity (i.e., the 'values') of the benchmark from which they depart. In short, the fact/value dichotomy prevails. The same methodological remarks apply to (2), though there has been much less work done on this axiom (chapter 4 and 5 of this dissertation are dedicated to it; indeed, as its name suggests, "description invariance" is the key implicit axiom preventing standard models to account for the uses of ordinary language in economic choices).

The following tension should be noted. On the one hand, reference-dependent preferences are often used to explain violations of the three invariance axioms (c.f., the references cited in this subsection). Whether or not loss aversion is involved in those explanations can be said to separate the behavioral economists (when it is involved) from the standard economists (when it isn't involved). Recall that we said at the beginning of this subsection that violations of (1), (2) and – though to a lesser extent – (3) were judged as signs of irrationality from behavioral economists and from the psychologists who demonstrated these violations. On the other hand, however, the notions of reference-dependence and loss aversion (which are used to explain those violations) are not necessarily considered as irrational. This is the case at least regarding the position of Tversky and Khaneman (1991) in their landmark formal contribution to reference-dependence and loss aversion *under certainty*, where they asked quite straightforwardly:

> "Is loss aversion irrational? This question raises a number of difficult normative issues. Questioning the values that decision makers assign to outcomes requires a criterion for the evaluation of preferences. [...] We conclude that there is no general answer to the question about the normative status of loss aversion or of other reference effects, but there is a principled way of examining the normative status of these effects in particular cases." (pp.1057-8)

In other words, the rationality or irrationality of loss aversion is highly situation-dependent. The tension mentioned above may be resolved quite straightforwardly: reference-dependent preferences and loss aversion are necessarily irrational at least when they lead to violations of (1) and (2), while there is still some room for arguments when they lead to violations of (3). However, it can be argued that the "criterion for the evaluation of preferences" or the "principled way of examining the normative" Tversky and Kahneman are talking about can be more nuanced than that. The next three subsections discuss three such nuances that are often made in behavioral economics and are not restricted to loss aversion or reference-dependency,

but applies to the broad issue of the normative dimension of models of individual behaviors.

Summing up, the factual observations of violations of procedure and description invariance from psychologist gave rise, through behavioral economics, to new theoretical conventions in economics, i.e., two new implicit axioms, and hence to new value judgments, i.e., violations of these axioms is irrational. This is different from Sen's work on context invariance and from *some* remarks made by Tversky on this axiom. Because violations of these three axioms go against the methodological convention of given preferences, it can be argued that Sen tends to argue for the rationality of endogenous preferences while behavioral economists *tend* to argue for their irrationality. The next subsections are meant to temperate this latter claim.

1.3.2 Kahneman on hedonic experience, or 'the primacy of time preferences'

The primacy of time preferences (over risk and social preferences) is implicit in the contributions that Kahneman has made (with various co-authors) on "experienced utility" from the 1990s onwards, and is usually associated with the revival of hedonic psychology (see esp. Kahneman 1994; Kahneman, Wakker and Sarin 1997; Kahneman and Sugden 2005; Kahneman and Thaler 2006). Experienced utility is indeed identified with the hedonic sensation of the consequence of a choice, and its maximization is taken to be a criteria of rationality that is empirically measurable. Using 'the primacy of time preferences' might be thought to be misleading here because the underlying model assumes no time preferences in the usual technical sense (i.e., with a discount rate and/or measuring the elasticity of intertemporal substitution, cf. next chapter). But I still want to use this expression to mark the fact that the concepts and notions constitutive of the model of experienced utility are inherently temporal, at least regarding decision makers' mental processes about time though *not* necessarily about their *behaviors* over time. Using Kahneman et al.'s (1997) terminology, the model of experienced utility consists in the (axiomatically characterized) relations among four (empirically measurable) concepts:

(i) "predicted utility": the beliefs about *future* hedonic experiences of "temporally extended outcomes", i.e., of a sequence of consequences of a choice;

(ii) "instant utility": the *present* hedonic experience of temporally extended outcomes

(iii) "remembered utility": the beliefs about *past* hedonic experiences of temporally extended outcomes

(iv) "total utility": the aggregation of instant utilities

As Kahneman et al.'s put it, "*[t]otal utility* is a normative concept" in their model. They discuss various experimental results showing that decision makers usually do not correctly anticipate or remember the hedonic experience of the consequences of their choices (i.e., they hold incorrect beliefs in (i) and (iii)). The hedonistic stance of the model implies that if instant utility is to be inferred from (i) by, e.g., observed choices, or from (iii) by, e.g., statements in questionnaires, then the biases in both should be *corrected* in (iv) the aggregation procedure. Or it implies that non-traditional techniques should be used to measure (ii) instant utility directly (see Kahneman and Sugden 2005, p.175 for a short discussion). In short, decision makers' expressions of preferences may not be taken directly as a 'given' in economic analysis, but they can be indirectly taken as such by the economist whence the appropriate corrections are made or the appropriate measurement techniques are used.

Notice that, within this framework, it is in principle possible to justify violations of procedure, description or context invariance when the procedure of elicitation and/or the description of the objects of choice and/or the context of choice impact the hedonic experience of a given choice. For instance, if you cannot help feeling better when paying a consumption good by cash when there is a 'credit card surcharge' by contrast with a 'cash discount' (e.g., because the former brings you the nice sensation to 'beat the system'), then a violation of description invariance can be justified as rational from Kahneman's experienced utility.[13]

Kahneman's (1994) methodological discussion of experienced utility is one of the rare instance where the work of Sen is discussed in (a contribution strongly related to) behavioral economics. It is first discussed to make a distinction: that experienced utility is a "*substantive* criteria of rational choice" akin to the ordinary language of "non-technical discourse", not a "*logical* criteria of rationality" as are conditions of internal consistency of choices in the formal language underlying "technical discussions" (Kahneman 1994, p.18, my emphases). Kahneman also discusses the work of Sen to highlight that while the latter emphasizes on the values ex-

[13]Further developments and empirical evidence about this reasoning are provided in chapter 4 and 5 of this dissertation.

pressed by decision makers and Kahneman on the inherent biases in these expressions, both are explicitly claiming a non-universal domain of application, and agree on each other' domain-dependent claims. In other words, the reliability of the decision maker's knowledge, or lack thereof, of his own values and preferences is domain specific (see esp. Kahneman and Thaler 2006 on that point). Furthermore, Sen's criticisms of hedonism (within his criticism of utilitarianism) are discussed to acknowledge that not only the knowledge of values are domain dependent, but also the values themselves, i.e., (instantaneous) hedonic experience is *not* taken to be a universal normative criterion applicable to all domains. One domain where hedonic experience is not a good criterion is where social and ethical values are involved, while other domains where it is thought to be a good domain is daily consumption goods (see Kahneman 1994, p.21; Kahneman and Sugden 2005, p.176). Finally, it is worth noting that Thaler explicitly endorses Kahneman's position (see Kahneman and Thaler 2006).[14]

Summing up, focusing on decision makers' relations to time allows the derivation of value judgments from hedonism eventually rationalizing factual violations of the theoretical conventions underlying that preferences are given. Though such reasoning can be justified within the framework of experienced utility, its valid application is acknowledged to be non-universal.

1.3.3 Tversky on reflexive reasoning, or 'the primacy of risk preferences'

The primacy of risk preferences (over time and social preferences) is implicit in some contributions made by Tversky on expected utility theory (Slovic and Tversky 1974; Tversky 1975; 1996). These contributions are embedded within the controversies on the normative status of the axiomatic structure of expected utility theory discussed in the next chapter. Quite generally, Tversky argues that to settle such controversies, the arguments of theorists should be presented to decision makers (who have or have not violated the axioms in experiments). And the tendency of the latter to be convinced by one side or the other of the controversies should itself be taken into account in the controversies. In other words, the inductive and deductive *reasoning*

[14] A comment by two behavioral economists on the contributions of Kahneman discussed in this subsection is worth quoting:

> "Some insightful economists such as Sen (1986) have for a long time harshly criticized the serious limits of analyzing human behavior by revealed preference alone. However, this had little effect on economics teaching and research as long as utility was thought to be unmeasurable." (Frey and Stutzer 2007, p.9)

capacities of decision makers are called for to settle some issues within the reasoning process of a scientific community. Again using the 'primacy of risk preferences' might be thought to be misleading because this experimental methodology is not directly about the measurement of risk preferences in the technical sense (i.e., the curvature of a utility function and/or of a weighting function, cf. next chapter). But I still want to use this expression to mark the fact that this methodology comes from and have been mainly used to study decision maker's behaviors under risk and *un*certainty. It nicely makes salient how the two main formal languages used in the search for *certainty* in science, i.e., probability theory and logic (viz. the axiomatic method) are constitutive of the issues raised in the literature on risk preferences, by contrast with the ones raised in the literature on social and times preferences. The contributions of Tversky under discussion here can be seen as a way of improving economists' and decision theorists' uses of those formal languages by studying their implications as seen by the economic agents.

The main point emphasized by Tversky throughout those contributions is that both the positive and normative dimensions of economic (or even human) behaviors are empirical matters. And the empirical inquiry into the normative dimension necessarily requires, on the one hand, a careful translation of the formal issues into the ordinary language that is shared by the economists and the economic agents, and, on the other hand, a careful reflexion from the decision maker during his reasoning process. These conditions can rationally influence changes in the initial ('given') preferences of the decision makers as well as in theorists' attitudes regarding the normative values of some axioms. In such situation, the final preferences expressed by a decision maker can be considered as endogenous to the interaction between him and the economist. To illustrate with the example from the previous subsection, imagine that a decision maker violates description invariance by paying with cash when there is a 'credit card surcharge' but with his credit card when there is a 'cash discount' (e.g., at different days or at different stores). Imagine further that the decision maker was being followed by two decision theorists disagreeing on the normative value of description invariance: one holds that the influence of descriptions on behaviors is irrational by following the requirement of standard models while the other holds that if descriptions influence experienced utility then this requirement should be relaxed. They decide to resolve their controversy by presenting their arguments to the decision maker and let the latter's evaluation of these arguments settle the controversy. If, after careful reflections, the

decision maker resolves the controversy in favor of experienced utility, then this warrants some weakening of this axiom in standard models.[15]

The role of *reasons* in such process has been further investigated by Tversky and his co-authors to show how "an analysis based on reasons may contribute to the standard quantitative approach based on the maximization of value" (Shafir, Simonson and Tversky 1993, p.33). As they point out, "[t]he reliance on reasons to explain experimental findings has been the hallmark of *social* psychological analyses" (p.34, my emphasis), which played a very small role (if at all) in the making of behavioral economics. Indeed, those contributions from Tversky are today seen to be at odds from behavioral economics. Recently, Eric Angner and Loewenstein (2012, p.667) compared the "incremental" approaches taken in most of behavioral economics' models who modify some behavioral assumptions of standard models (see next chapter) to the reason-based contribution from Tversky and his co-authors. They claim the latter to be among the "radical approaches [that] try to improve the predictive power and explanatory adequacy of current theory *by starting from scratch*" (ibid, my emphasis). So the message from Tversky and his co-authors that there are important complementarities to be developed between their reason-based approach and the standard model does not seem to have come across. It will be argued in the next section that much of Thaler's work in normative economics share nontrivial similarities with a reason-based approach.[16]

How does this fit within the comparison between Sen and behavioral economics? Reason-based approaches are well in line with Sen's overall approach to rationality as "reasoned scrutiny" (2002, chap. 1, sect. 13). Roughly, Sen's reasoned scrutiny aims at articulating the different uses of rationality in economics, and argues that the role of *reasons* in the economists' work and the economic agents' decisions is necessarily central in that task. Sen himself acknowledges that his broad interpretation of rationality and the openness of the approach he preaches are subject to a number of criticisms, to which he offers some replies. Notably, to the objection that his approach provides no clear criteria regarding what counts as rational or not in a given situation, Sen replies that "while there is scope for much work on the type of criteria that may

[15]This scenario is a very rough idealization of what happened to the independence axiom of expected utility theory (cf. next chapter) and what may be happening to description invariance (cf. chapters 4-5).

[16]For recent developments of the reason-based approach, see Dietrich and List (2013; 2015) and Lecouteux (2015).

be used, [...] maximizing the domain of applicability, irrespective of the quality or cogency of that application, is not necessarily a great virtue for a critical discipline" (2002, chap.1, p.49). It can be argued that the experimental methodology underlying Tversky's contributions discussed above allows for a disciplined yet situation-dependent way of assessing criteria of rationality. Furthermore the exchanges of reasons between economic theorists and economic agents that takes place in these experiments is in line with the openness of Sen's reasoned scrutiny.

Summing up, the primacy of risk preferences in the work of Tversky refers to the methodology used primarily in his experiments about risk preferences. The general idea is to let the reflexive reasoning of decision makers be the arbiter of theoretical disagreements on the normative dimension of models of individual behaviors. It can be extended straightforwardly to any dimension of economic behaviors where there is such theoretical disagreement and is well in line with Sen's approach to rationality as reasoned scrutiny. The next subsection provides further characterization to Sen's reasoned scrutiny.

1.3.4 Sen on reasoned scrutiny, or 'the primacy of social preferences'

In this subsection, I wish to argue that there is a primacy of social preferences in Sen's conception of rationality as reasoned scrutiny. Indeed, it is not a coincidence if the first contribution where Sen (2002, chap.1) characterizes what he means by reasoned scrutiny is also the first contribution where he discusses behavioral economics at some length *and* that this discussion focuses on accounts of social preferences within the latter (notably Thaler's). It can be argued that the very introduction of 'reasoned scrutiny' by Sen (2002, chap. 1, sect. 10) is meant to articulate most of the conceptual apparatus he developed in earlier criticisms of standard models of individual behaviors. More precisely, under this interpretation, reasoned scrutiny makes a conceptual connection between, on the one hand, his threefold characterization of the notion of self-interest, and, on the other hand, his distinction between sympathy and commitment. This subsection explains this conceptual connection and discusses one of the rare instance of experimental work inspired from Sen's conceptual apparatus.

Besides the revealed preferences tradition imposing consistency conditions of rationality on choice functions, other traditions (esp. the Chicago School) imposes as a requirement of rationality on behaviors that they can be *interpreted* as the maximization of the decision maker's

'self-interest'. By contrast with the former tradition, controversies around self-interest have been less concerned with technical issues within a formal language than with the interpretational issues arising from the use of ordinary language. Hence for the purpose of "distinguishing between the distinct [interpretational] issues involved in the contemporary debates on the role of self-interest in rationality" (2002, chap.1, p.33) Sen has characterized "three requirements" usually "imposed in the traditional models" (ibid, p.34):

> "*Self-centered welfare*: A person's welfare depends only on her own consumption and other features of the richness of her life (without any sympathy or antipathy towards others, and without any procedural concern).
> *Self-welfare goal*: A person's only goal is to maximize her own welfare.
> *Self-goal choice*: A person's choices must be based entirely on the pursuit of her own goals" (ibid, pp.33-4)

Sen argues that interpretations of self-interest through standard models of individual behaviors typically take any of these three forms or any combination of them. On Sen's account, these three requirements usefully capture "three different aspects of the "self"" (ibid). However, none of them and no possible combination of them can capture a fourth aspect of the self which, he argues, is crucial to characterize human rationality, namely the capacity "to do self-scrutiny and reasoning" (ibid). Roughly, reasoned scrutiny refers to the decision maker's ability to critically examines his own welfare or goal to eventually change them. In other words, it refers to a conscious process of endogenous changes in preferences, during which they are therefore not a 'given' for the decision makers. This is indeed the reasoning driving Sen's explanations of violations of context invariance (see Sen 2002, chaps.3-4).

With respect to this set of distinctions, Sen's further distinction between sympathy and commitment marks two "possible foundations for other-regarding behavior" (p.35), i.e., two ways interpretations from the three requirements or their combinations may be criticized. Sympathy is the uncontroversial part of Sen's distinction: "one person's welfare being affected by the state of others" (2002, ibid); or, put in words appealing to a behavioral economist, "[w]hen a person's [...] well-being is psychologically dependent on someone else's welfare" (1977b, p.327). It necessarily violates the first requirement but not necessarily the two others. We shall see in the next chapter that behavioral economics accounts of social preferences can be interpreted as different models of sympathy but not of commitment. Indeed, the latter, which is "concerned

with breaking the tight link between individual welfare (with or without sympathy) and the choice of action" (2002, ibid), is the controversial part of Sen's distinction. By contrast with sympathy, commitment does not necessarily break any one of the these three requirements. But it can, and its conceptual usefulness lies precisely in potential violations of self-welfare goal and *especially* self-goal choice. The connection with reasoned scrutiny arises at this juncture, namely in explaining violations of self-goal choice, which imply an action that is motivated by something else than one's own goals, i.e., by other people's goals or by any other *reasons* not necessary for the pursuit of others' or one's own goals, e.g., *some* "moral or social or political reasons" (Sen 2002, p.36). Hence the central explanatory feature of reasoned scrutiny is its potential production of such reasons to motivate choices.[17]

It is often argued that interpretational issues are looming large in Sen's conceptual apparatus (see Peter and Schmid 2007). In what sense can either someone else's goal or the moral or social or political reasons be any different from my own goals when they motivate my actions? How can it be that I can fulfill goals without increasing my welfare? Besides philosophical arguments, there have been no experimental investigations of Sen's conceptual apparatus to answer such questions. Partial answers can however be found in Allan Shiell and Bonnie Rush's (2003) empirical study in the health domain. Though it is not an experiment strictly speaking (there is no control condition or even different conditions to be compared), the design of their survey contains questions that are akin to the ones posed in lab experiments. There are two steps. First, decision makers are asked their willingness to pay for two vaccination programs: one program benefits only to the decision maker while the other program does not benefit the decision maker but benefits other people among the poorest 10% of the population. Second, structured interviews with these decision makers are conducted and recorded, from which they then extract verbatims that allow to disentangle self-centered, sympathetic, and committed reasons of decision makers. For instance, the three following verbatims illustrate each of these types of reasons (pp.656-5):[18]

[17]The important link between the expressions of commitment through reasoned scrutiny in ordinary language and the uses of partial ordering in Sen's formalization of context-dependent preferences is worth mentioning but will not be discussed in this dissertation (see Sen 2002, part I).

[18]To be more precise on their methodology, their interviews contain notably two types of closed questions: "'how difficult did you find the exercise?' and 'does your reported willingness to pay reflect the value that you place on the vaccine?"' (p.652). Shiell and Rush's quantitative data consist in the willingness to pay, and in a dichotomization of answers to the two previous questions into "'some or no difficulty' or 'difficult or very difficult'

S*elf-centeredness* from a decision maker willing to pay $0 for the others:

> "There is no direct benefit to me. I hate to sound cold, but if there is no direct effect then there is no real value to me."

Sympathy from a decision maker willing to pay $60 for the others:

> "(I was) trying to assess whether the increase in costs to myself would compare with the value that I would place on extra coverage for other people."

Commitment from a decision maker willing to pay $10 for the others:

> "the benefits of the program are not worth anything to me because I would not be receiving anything from the vaccination . . . . I would feel comfortable paying $10 to have someone else have a vaccination (but) I have nothing to gain from it. It doesn't affect me in any way."

Two points that are not often stressed (though they are not denied either) in the conceptual discussions of commitment can be emphasized here. First, commitment does not imply irrational degrees of generosity, as illustrated by the committed decision maker who gives less than the sympathetic one. Second, the observability of committed behaviors necessitates a communicative structure between the observer (e.g., the economist) and the decision maker that goes beyond the elicitation of choices as traditionally conceived. That does not mean that commitment cannot be observed in choice experiments, but that their observability necessitate the exploitation of extra data from choice data. Or in Sen's words: "empirical evidence for this cannot be sought in the mere observation of actual choices, and must involve other sources of information, including introspection and *discussion*" (1977b, p.342, my emphasis).

Summing up, by contrast with the primacy of time and risk preference in the work of Kahneman and Tversky (respectively) the primacy of social preferences in Sen's rationality as reasoned scrutiny is much closer to the technical meaning of social preferences in economics (outside of the 'aggregation' meaning in social choice theory, however) as characterizing departures from self-interest. Reasoned scrutiny however emphasize the endogenous process of preference revision through critical self-examination during which preferences are not taken as a 'given' by

and 'yes' or 'no' respectively" (p.653). Their qualitative data consist in the recording of the interviews, from which they extract the verbatims; the ones presented here are extracted from answers to the second question, i.e., 'does your reported willingness to pay reflect the value that you place on the vaccine?'

decision makers. It is through this process that Sen's controversial 'commitments' emerges, which we saw are observable though not directly as choice behaviors but through the uses of ordinary language by the decision makers.

Conclusion

This section developed some points of convergence between the criticisms from Sen and from the behavioral economics of Thaler to the methodological convention of taking preferences as 'given' in standard economic analysis. Violations of three theoretical conventions underlying standard models – procedure, description and context invariance – are the main phenomena that the psychologists who inspired behavioral economics have demonstrated to highlight the limits, from a positive perspective, of this methodological convention. Both these psychologists and behavioral economists usually derive value judgments from standard models to argue that the factual behaviors are irrational and the theoretical conventions are rational. However, looking for arguments that would go against this reasoning *within* the work of these psychologists, we found some from hedonic experience in the work of Kahneman and from reflective reasoning in the work of Tversky. We argued that none of them are incompatible with Sen's *normative* stance against the methodological convention of given preferences from reasoned scrutiny. On this matter, there is more than an absence of incompatibility between Tverksy's and Sen's works; some fruitful mutual developments could be made. Indeed, Tversky's work can be seen as an experimental implementation of Sen's reasoned scrutiny where a communicative structure is designed for the purpose of resolving theoretical controversies between decision theorists through the reasoning of decision makers. However, for the fruitfulness of such mutual developments to be maximal, data about the uses of ordinary language should be collected *besides* choice data, in order to make the reasons behind choices empirically observable.

It was also mentioned that Thaler's work in normative economics bears some resemblance to the work of Tversky discussed here. The next section will develop this remark as it scrutinizes, by contrast with this one, the strongest points of disagreements between Sen and behavioral economics, namely about the articulation between positive and normative economics.

1.4 Strong disagreements over the articulation of positive and normative economics

The goal of this section is to conclude the comparison between Sen and the behavioral economics of Thaler on the issue about which they disagree the most, namely the articulation between positive and normative economics. We shall focus on how the entanglement thesis highlights that the normative dimensions of models of economic behaviors is partly influenced by some features of the articulation between positive and normative economics (1.4.1). Thaler's position is presented with an emphasis on its proximity with Tversky's work from the previous section and on the proximity between what he calls 'choice architects' and what we called 'decision modeler' (1.4.2). His position is then contrasted with Sen's (1.4.2). Finally, the strong disagreements are situated within a rough sketch of the current methodological reflections on the articulation between positive and normative economics (1.4.4).[19]

1.4.1 The normativity of rationality that comes from normative economics

For starters, some clarifications regarding 'normative economics' and 'welfare economics' are in order. As remarked by Mongin (2006a), the historical developments of different areas of normative economics in the postwar era (notably social choice theory and public economics) made it so that "[f]ew works with the title "welfare economics" were published beyond the 1960s" (fn.10 p.24). The expression 'welfare economics' is however not out of date because, as Antoinette Baujard shows, it is broadly understood as "the economic study of the definition and the measure of the social welfare; it offers the theoretical framework used in public economics to help collective decision making, to design public policies, and to make social evaluations" (forthcoming, first page). On that account, it is nearly synonymous with 'normative economics',

[19]This section is the only place in this dissertation where the positive/normative issue understood as the characterization of subareas of economics into 'normative economics' and 'positive economics' will be scrutinized. Since the problems and references discussed here are very partial, the interested reader can find the adequate bibliography in Hands' (2012b, sect.2) chapter to which one could add Brochier (1995) and Su and Colander (2013). The latter reference is specifically concerned with the entanglement thesis and assesses a debate between Putnam and Walsh *versus* Partha Dasgupta (2005; 2007; 2009). This debate will not be discussed here because it would take us too far of the implications of the entanglement thesis *for behavioral economics*. However, a good illustration of the heterogeneity of what counts as standard welfare economics among self-identified standard economists can be drawn by comparing Dasgupta's position with the positions in Caplin and Schotter's (2008a). As for the recent debates around normative economics triggered by behavioral economics, see Lecouteux (2015) and the references therein.

especially in so far as

> "it is notoriously hard to say what exactly normative economics is about – welfare or choice, value judgments or the study of value judgments, economic policy or armchair evaluation. Economic methodologists or theorists have provided grand statements on how normative economics should be separated from positive economics and applied economics" (Mongin 2006a, pp.19-20)

One of the least controversial way of viewing the matter (which we shall endorse in this dissertation) is this: "the task of normative economics is to investigate methods and criteria for evaluating the relative desirability of economic states of affairs" (ibid, p.20).

The entanglement thesis can be seen as one of those "grand statements on how normative economics should be separated from positive economics" (ibid), stating that it *not* be *separated* but articulated differently. Sen opens the second chapter of *On Ethics and Economics* by a criticism of the "one-way relationship by which endings of predictive economics [i.e., positive economics] are allowed to influence welfare economic analysis, but welfare economic ideas are not allowed to influence predictive economics" (Sen 1987, p.29). Hence "[w]elfare economics has been something like an economic equivalent of the 'black hole' – things can get into it, but nothing can escape from it" (ibid). Despite this "one-way relationship", Walsh (1996) points out *an implicit methodological impact of normative economics on positive economics* that shows how any kind of separation thesis is untenable given the theoretical structure of economics. One of the main inputs from positive economics into normative economics is general equilibrium theory, to which the application of the normative criterion of Pareto-optimality yielded the two fundamental theorems of welfare economics. One way of putting these two theorems in ordinary and non-technical language is as follows. Given some assumptions about production and *some assumptions about consumption, i.e., about consumer's preferences*:

1st Theorem: every competitive general equilibria in an economy, i.e., every states of the economy where market interactions determine the prices at which all production is consumed, will be Pareto-optimal, i.e., there are no other allocations of the production to the consumers that will make all consumers better-off;

2nd Theorem: every such Pareto-optimal states of an economy can be attained through competitive mechanism, i.e., by prices determined through market interactions instead

of State decisions, if consumers' initial budgets are distributed in certain ways.

Walsh (1996) argues at length that one retroactive methodological consequence of those results from normative economics on positive economics is the strengthening of the normative dimensions of models of economic behaviors used in consumer choice theory. This is so because the axioms on which these models are constructed *are* the assumptions on consumer's preferences *needed* for the proofs of the theorems. Hence the formal properties of standard models of economic behaviors are an especially nice form of rationality *because* of the role they play "at the heart of elaborate proofs" of "equilibrium allocations, [which are] the rational dispositions of goods resulting from the rational choices of every individual agent in such models" (Walsh 1996, p.161). Therefore, a part of the normativity in standard models of economic behaviors comes from the role they play at another theoretical level (i.e., general equilibrium theory), in making possible that *the economy* reaches an *optimal* or *efficient* state. To put it crudely, one reason why economic agents *ought* to behave according to the standard models of behaviors is that, if they don't, they and every other economic agents will not be better-off and the economy will not be efficient.

Furthermore, the standard *interpretation* (i.e., not only the mere uses) of models of individual behaviors in those results further explains parts of the normativity of the methodological convention of *given* preferences in positive economics. Indeed, if there are any meanings to those two theorems being called 'fundamental theorems of *welfare* economics', the 'better-off' relation in their statements has to be interpreted in terms of the consumers' welfare. A standard theoretical move here consists simply in *defining* individual welfare as individual choice, preference or utility, with the standard conditions ensuring their equivalence, i.e. $C(.) \Leftrightarrow \succsim \Leftrightarrow U(.)$ discussed in the first section. This theoretical move, though standard, is often strongly disputed (see esp. Sen 1973, sect. VI; Mongin 2006a, sect. 8). Roughly, just as there are problems in positive economics because, e.g., one can have plenty of *reasons* not to prefer what he chooses, the problems in normative economics are that, e.g., one can have plenty of reasons to derive some welfare from what he does not prefer, to choose to decrease his welfare, etc. (see Sen 1987, chap.2; Hausman 2012, part. II). What matters for the present purpose, however, are less the problems underlying that theoretical move than *its justification by consumer sovereignty*, maybe the most widely shared value judgment among standard economists (see Juille 2015). As Blaug

explains:

> "the modern doctrine of Pareto optimality rests among other things, on the fundamental postulate of consumer sovereignty – only self-chosen preferences count as yardsticks of welfare or, in popular parlance, an individual is the best judge of his or her welfare – and it has long been argued that consumer sovereignty is a value judgment *par excellence*, implying that Paretian welfare economics is fundamentally normative. However, [some] have argued instead that the theorems of Paretian welfare economics are theorems in positive economics; on this view, the assumption of consumer sovereignty is not a value judgment but simply the assertion of the axiom that individual preferences are to be taken as given for purposes of assessing a potential Pareto improvement, without endorsing or approving of these preferences" (1998, p.373)

If the discussion of given preferences in the previous section is correct, then the argument in the last sentence of the quote is flawed. Working with given preferences does not *only* consists in not endorsing or approving their contents, but *also* consists in taking them to be exogenous to the economic situation under study. In other words, the potential Pareto improvement will just satisfy preferences without changing them. For instance, if the government decreases the price of candies, it should just satisfy my preference for candies over vegetables, not change it for the reverse because, say, I automatically disprefer any consumption goods subject to state intervention. In any case, what is at stake here are value judgments. The difference is that it is an ethical value judgment (consumer sovereignty) in the normative economics interpretation – "an individual is the best judge of his or her welfare" (ibid) –, and an epistemic or methodological value judgment (given preferences) in the positive economics interpretation – 'individual preferences are exogenous variables in the study of an economic situation'. In both cases the result is however the same: economists cannot or should not judge economic agents' judgments, i.e., economists ought not to question the economic agent's *reasons* for their choices. And that, it can be argued, goes a long way towards explaining the reluctance of economists for reason-based models of choices (at least up until recently) we observed in the previous section.[20]

[20] Cowen (1993, sect. 2) discusses more fully the problem posed in this paragraph with references to other economists who stated it. I thank Tom Juille for this reference.

1.4.2 Behavioral economics: disappearance of *theoretical* normative economics?

In the previous section, we also observed that contemporary behavioral economists shared such reluctance, contrary to the work of Tversky and Sen. I would now like to suggest that although Thaler's stance in normative economics is not incompatible with a reason-based approach, his closeness with the standard approach on Pareto-optimality and consumer sovereignty may explain why reason-based approaches are unpopular in behavioral economics.

The book *Nudge* (Thaler and Sunstein 2008) has raised a host of debates around the new form of paternalistic interventions in the economy that Thaler and Cass Sunstein defend under an approach they label 'libertarian paternalism'. One methodological point of their approach is that they "do not always equate revealed preference with welfare" (Thaler and Sunstein 2003, p.175), and economic agent's welfare is not supposed to be determined by the economist but by the economic agent's own reflective reasoning. Thaler colorfully puts the point and the issues around it as follows:

> "a point that critics of our book seem incapable of getting: we have no interest in telling people what to do. We want to help them achieve their *own* goals. Readers who manage to reach the fifth page of *Nudge* find that we define our objective as trying to "influence choices in a way that will make choosers better off, *as judged by themselves*." The italics are in the original but perhaps we should have also used bold and a large font, given the number of times we have been accused of thinking that we know what is best for everyone. Yes, it is true that we think that most people would like to have a comfortable retirement, but we want to leave that choice up for them. We just want to reduce what people would themselves call errors" (2015, p.326)

But how does the economist know what the economic agent judges to be his own goals, if his observable choices don't reveal it? One usual answer is that observable conflicting choices, e.g., buying cigarettes to smoke *and* buying drugs to stop smoking cigarettes, provide enough information to infer that the economic agent's own goal is to stop smoking. Another usual answer is that the economic agent just expresses his goal straightforwardly, as in a famous questionnaire study often referenced by behavioral economists where the majority of Americans state that they don't save as much money as they would like to (though they could). Notice that in both cases, it is the *p*sychology and ordinary language shared by the economist and the

economic agent that allows the former's inferences. More controversially, another usual answer is that economic agent's own goals are the ones corresponding to the choices they would made "if they had complete information, unlimited cognitive abilities, and no lack of willpower" (Thaler and Sunstein 2003, p.175). In most cases, however, behavioral economists limit themselves to situations where the agents' own goals seem "profoundly plausible and of great policy relevance" (O'Donoghue and Rabin 2003, p.191). That is, they "skirt the metaquestion of what it means for a person's behavior to be in their "best interest" [...] [and] simply assume that if an individual expresses a consistent desire to achieve a certain goal, such as losing weight, taking medications or saving money, it is relatively unobjectionable to help them achieve that goal in a fashion that does not restrict their ultimate freedom of choice" (Loewenstein et al. 2013, p.362).

Such an insistence on not restricting freedom of choice is widespread in the great majority of behavioral economists' contributions to normative economics. It is in fact necessary for another requirement of their approach, namely that the interventions that increase the welfare of those who made mistakes do not decrease the welfare of those who don't. In other words, not restricting freedom of choice ensures that the existence of conflicts among the goals of economic agents, e.g., those whose own goals are to save more money and those whose own goals are to save less money, will not decrease the welfare of those who are not targeted by the intervention (see esp. Camerer et al. 2003). For instance a traditional policy intervention on some rate of interests related to savings impacts (positively or negatively) the welfare of all economic agents indiscriminately, while in behavioral economists' proposals of different saving plans, economic agents are still free to choose to save as much as they want after the intervention (e.g., by refusing the plan). Such insistence on freedom of choice for the sake of improving the welfare of some without decreasing the welfare of others is *a clear (but not formally explicit) commitment to the standard defense of the criterion of Pareto-optimality through (a slightly altered version of) consumer sovereignty.*

If behavioral economics is methodologically very close to standard normative economics on that point, they are not close to it on the point that "paternalism is unavoidable: behavior is shaped by people's environments, and environments have to be structured in some way; *there is no neutral way to structure an environment*" (Loewenstein et al. 2013, p.362, my emphasis). As Thaler and his co-authors have recently put it:

> "everything matters. Tiny details, from the color of an alert lamp to the size of the font can influence choices. [...] [S]ince everything matters, it is important for those who design choice environments whom Thaler and Sunstein (2008) call "choice architects," to take human factors into account" (Benartzi et al. 2013, p.245)

Indeed, all the discussions in *Nudge* are centered around the notion of a choice architect, who, according to Thaler and Sunstein should better know the empirical evidence about human behaviors gathered by the behavioral sciences (including behavioral economics) for their choice architecture activities to be successful (see also Thaler et al. 2013). That is what is meant in the previous quote by taking the "human factors into account". All the features of the role of a choice architect in the activity of choice architecture discussed by Thaler are well in line with the role of the decision modeler in the communicative structure of economic choices discussed two sections ago. Especially worth noting is Thaler's emphasis on how the *important* economic choices invariably involve choice architects' uses of ordinary language, e.g., describing options and consequences in the drawing of a contract, be it for health care, employment, sales, etc.. That "[w]hen dealing with Humans, words matter" is carefully stressed by Thaler (2015, end of chap.32). As he puts it while describing one of his policy making experience: "[a]ll we had to do was fiddle with the wording of a letter that would be sent to taxpayer anyway [...], fine-tuning the letters could potentially save millions of pounds" (chap.33; the whole chapter emphasizes the importance of *different* impacts of *same* propositions expressed using ordinary language *differently*).

One development from Thaler and Sunstein to our communicative structure of choices is their discussions of architects are themselves making *choices*, whether unconsciously or not. That is, the consequences of their actions *are* a certain presentation of a choice set to a decision maker, *while other presentations would have been possible.* That matters because even if a certain presentation is chosen for no specific reason, decision makers will tend to infer (incorrectly) such a reason. For instance, when one option is put as a default and decision makers infer (correctly or incorrectly) that the fact that this option rather than another is the default is an implicit endorsement of this option by the choice architect (Thaler and Sunstein 2008, p.35). That is another way of saying that "there is no such thing as a "neutral" design" (p.3), or in the terms used two sections ago, that decision makers are not facing the brute force of a purposeless Nature. One tension in their account should however be noted. If decision makers are necessarily

influenced by choice architects, then decision makers need to be aware of this influence when they are expressing what makes them better off 'as judged by themselves'. This seems to require the exercise of reasoned scrutiny as discussed in the previous section, which itself requires the expressions of the reasons that motivate what decision makers would like to do. Following the arguments from the previous section, without a way of expressing those reasons through ordinary language, the observing behavioral economists might miss some commitments that are important for decision makers. The lack of popularity of reason-based models in behavioral economics seems problematic regarding these issues.

Another development from Thaler's choice architects may be his emphasis on how knowledge from the social and behavioral sciences can impact the choice architect's or the decision modeler's activities. By contrast, one difference between both account may be that Thaler's choice architect is never identified with an economist doing positive economics, e.g., conducting a lab experiment. However, recall that one of the main purpose of making the distinction between decision modeler and decision maker was to account for such existing communicative interactions between economists and economic agents. If this is indeed the difference between Thaler's choice architect and our decision modeler, then the latter is a bit more general than the former, i.e., by including the possibility of identification with the activity of economists doing positive economics.

To conclude, it should be noted that contrary to existing theoretical contributions of behavioral economists on models of individual behaviors, much of their contributions to normative economics are not strictly speaking theoretical. That is at least the case for Thaler's contributions, which are rather straightforward policy proposals based on behavioral evidence (from the lab or from the field), that are aiming to change some targeted individual behaviors. Neither theoretical models of individual behaviors explaining the behavioral evidence through *P*sychology, nor (and even less) potential formalized frameworks of normative economics play a crucial role *in the justification* of the proposals. What plays a crucial role are the existing behavioral evidence, not only about individuals behaving in certain ways to be changed, but (and especially) about the effectiveness of the proposed policy to change those behaviors in the advocated way. That is all explicitly part of the so-called (behavioral) 'evidence-based policy' movement (Shafir 2013; Thaler 2015, part. VIII; for a reflexive perspective see Favereau 2014, pp.96-106, 268-305). On

this account, normative economics is just an application of the empirical findings of positive economics. The economist has nothing specifically important to say about the values at play either in the policy proposal at stake or in the underlying behaviors of the economic agents. More than a clear separation between normative economics and positive economics, if normative economics is supposed to have something of scientific interests to say about such values, then it seems that normative economics has disappeared.

1.4.3 Sen: the disappearance of a *separate* normative economics

On this latter point, the contrast with Sen's articulation of positive economics and normative economics, at the heart of the entanglement thesis, is radical. As remarked by Mongin (2006a, sect.8), Sen is one of the main author marking a rupture between past normative economics and contemporary normative economics, a rupture characterized mainly by a dialogue with ethics, moral and political philosophy (maybe the evidence policy movement, not discussed by Mongin, is an even newer rupture with no such dialogue). One methodological specificity of Sen's position in that rupture is to make such dialogue not only useful for normative economics, but for positive economics as well; and also for ethics, moral and political philosophy themselves for that matter (see Sen 1987). Just as these interdisciplinary exchanges provides insights about the plurality of motives underlying individual choices (relevant for accurately characterizing them in positive economics, as we saw above), they also provides potential alternative criteria to Pareto-optimality for the evaluation of economic states of affairs. It is not so much Pareto-optimality by itself that Sen criticizes throughout his work in normative economics, but mainly the lack of pluralism of standard economists who do not seriously consider other alternative criteria better suited in some situations (based on rights, or equity, or fairness, among other well-discussed criteria in ethics, moral and political philosophy).

Contrary to behavioral economics' contributions to normative economics, most of Sen's are theoretical in the sense of providing an *articulated framework* of concepts and measurement procedures where the explicit role of (the economists' *and* the economic agents') value judgments looms large. The 'capability approach' is the label of the framework for normative economics in which Sen articulates the pluralism mentioned above. That approach offers a "freedom-based" middle ground between the two main approaches used to evaluate economic states of affairs (Sen

2009, p.231). On the one hand, there is the "utility-based" approach from welfare economics, often taken to be 'subjective' (ibid). On the other hand, there is the "resource-based" approach based on income, wealth, GDP and the like, often taken to be 'objective' (ibid). One salient characteristic of the capability approach is to evaluate the quality of life inherent to an economic situation by putting some weights on, say, education, without reducing it to the utility economic agents eventually derive from education, or to the economic consequences of education in terms of growth for instance. That implies enriching the informational basis of normative economics so as to include the content of the agents' preferences and the reasons they may have to value some objects of choice intrinsically, but not necessarily instrumentally, as in terms of utility from the agent's perspective or in terms of GDP from the economist's perspective.[21]

Sophie Pellé remarks that "by proposing to evaluate welfare in terms of individual 'capability', the position defended by Sen distances itself from consumer sovereignty in so far as it conceives welfare in terms of elements that should be satisfied, whatever individuals are thinking" (2009, p.151, my translation). The elements that should be satisfied are not predetermined in Sen's framework, they must emerge through – or as Sen often puts it, they must 'survive' – the economic agent's reasoned scrutiny, as well as some collective (maybe public, and possibly political) process of reasoned scrutiny – especially in the assessment of economic states of affairs (instead of individual decision making). Although Sen is much more critical than behavioral economists on the centrality of Pareto-optimality, his position on consumer sovereignty is not so different than theirs. Both so to speak 'weaken' consumer sovereignty based on mere choice towards consumer sovereignty based on a form of reflexive reasoning from the consumer (whatever they are thinking, they ought to undertake such reflexive reasoning). As Sen puts it, "the idea of freedom also respects our being free to determine what we want, what we value and ultimately what we decide to choose." (2009, p.232). Where Sen's weakening is stronger is in the inclusion of a collective process in the justification of the uses of consumer sovereignty as a criteria for the evaluation of economic states of affairs (ibid, pp.241-247). That means at least the economic agent's exercise of "comparative judgments" (p.243), the confrontation of his own

[21] This is a very partial account of Sen's capability approach. Besides Sen's own work (the references of which are to be found in the following ones), lengthier accounts that are in line with the entanglement thesis have been offered by Pellé (2009), Gilardone (2010), Nuno Martins (2013) and Davis (2011, chap.10) among others and besides the papers in Walsh and Putnam's (2011) volume.

reasons with other people's reasons, for his reasoning process to qualify as reasoned scrutiny. Notice that it is perfectly possible in such reasoning process that the economic agent confronts his reason to the economists', and whether the latter is the decision modeler (or choice architect) in such process becomes arguably an important issue.

As Putnam and Walsh put it while commenting on Sen's methodological position about the articulation between positive and normative economics, "the logical entailments of entanglement render the concept of a separate 'welfare' economics meaningless" (2009, p.291). Indeed, methodologically speaking, in scrutinizing the "triple entanglement: of fact, convention and value" (Sen 2005, p.112), the activity of describing economic states of affairs and of evaluating them can still be relevantly distinguish, but not relevantly separated in the sense of making them independent. Some descriptions are more accurate when made through an enriched evaluative framework that includes considerations of social justice, e.g., a woman paid less than a man for the same job is more accurately described as 'suffering from gender inequality' than as 'revealing a preference for a smaller wage'. Conversly (but more trivially), some evaluations are more accurate when made through an enriched description of the economic state of affair to be evaluated, e.g., an evaluation of the previous example is more accurate with some descriptive accounts of woman-man productivity differentials and of woman's bargaining power in wage negotiation. That might seem disturbingly obvious at first sight. But if such claim is worth stating and even defending at all, as Putnam, Sen and Walsh have done, it is because it goes against the dominant positions in the current methodological reflections on the articulation between normative economics and positive economics.

1.4.4 A picture of the contemporary methodological reflections

Mongin (2006c) provides a thorough systematization of the different positions taken by various economists at various times on the articulation between positive and normative economics. He characterizes these positions by three mutually exclusive thesis that they may hold, to which he adds a fourth one he defends. However, Mongin discusses neither the debates around behavioral economics nor the entanglement thesis. Hence, to conclude this section, I would like to enrich the picture he offers by, on the one hand, situating in it the debates around behavioral economics, and, on the other hand, adding a fifth thesis corresponding to the entanglement thesis. It may

	The neutrality thesis		The non-neutrality thesis	
	Strong version	Weak version	Weak version	Strong version
Economists...	... should never make value judgments	... can make a clear set of value judgments **separable from factual and analytical judgments**	... can make value judgments, but some are **not separable from factual judgments**	... cannot avoid making value judgments
Normative economics...	... is not scientifically legitimate, only positive economics is	... is scientifically legitimate, but **separated from positive economics**	... is scientifically legitimate, but **not separated from positive economics**	... is scientifically legitimate, the pretension to do purely positive economics is not
Current situation	Gul and Pesendorfer (2008)	Behavioral economists; *Foundations* (2008)	Mongin (2006c)	
			Moderate version	
			The entanglement thesis: some value judgments are indistinguishable from facutal ***and analytical*** judgments	

Table 1.1: Contemporary positions on the articulation of positive and normative economics

be more convenient to first have an overall illustration of this enrichment as provided in Table 1.1, and then discuss the meaning of its content.

On Mongin's account, both versions of the neutrality theses have been discussed and disputed within standard economics, while the strong version of the non-neutrality thesis has come from heterodox economics and other social sciences outside of economics, where it also applies.

Mongin's clarification holds pretty well for the contemporary literature around behavioral economics on the matter, at least as posed in the *Foundations of Positive and Normative Economics* (Caplin and Schotter 2008a). On the one hand, Gul and Pesendorfer's (2008) defense of standard economics is based on the strong version of the neutrality thesis. More precisely, one of Mongin's remark (2006c, p.259) characterizes particularly well Gul and Pesendorfer's position: defendants of the strong neutrality thesis either see welfare economics as not being economics but ethics (this is how Gul and Pesendorfer see welfare economics when it formulates policy recommendations) *or* as positive economics because its formalization through mathematics and logic makes it value-free (this is how they see the relevance of welfare economics, i.e., as raising

questions for positive economics).[22]

On the other hand, behavioral economists (Köszegi and Rabin 2008b; Loewenstein and Hainsley 2008) and standard economists strongly sympathetic to behavioral economics – or 'formerly-standard-but-recently-turned-behavioral economists' – (Bernheim and Rangel 2008) defend variants of the weak version of the neutrality thesis. They also follow the traditional strategy associated to it by Mongin, namely to argue that judging economic states of affairs in terms of Pareto-optimality for the sake of making evaluation, recommendation or prescription is scientifically legitimate. One small difference is that Pareto-optimality is weakened in the discussion around behavioral economics as explained above, or that some other criteria are so weakened. Especially relevant for the characterization of the current situation is Bernheim and Rangel's (2008; 2009) weakening, which does take the form of a theoretical and formalized framework to conduct welfare analysis, unlike behavioral economists' work in normative economics as explained above.

Nothing quite like Mongin's (2006c) own weak non-neutrality thesis occurs in the *Foundations*. That is not so surprising because the thesis Mongin defends is philosophically sophisticated in bringing altogether arguments from the philosophy of economics, ethics and moral philosophy, along with the philosophy of language, which *might* seem *at first sight* too exotic to be discussed in such volume. Parts of Mongin's (2006c) arguments (regarding thick predicates, see the general introduction of this dissertation) are the same ones that are at the core of the entanglement thesis, notably regarding the impossibility of an absolute dichotomy (i.e., the non-necessary independence) between judgments of values and judgments of facts. That is the core of Mongin's thesis (as mentioned in the table above). Recall from the general introduction of this dissertation that the entanglement thesis borrows further arguments from yet other areas of philosophy – of science , of mathematics and of logic – arguing for the impossibility of an absolute dichotomy, in science, between judgments of facts and judgments of analyticity

[22] Caplin and Schotter's (2008b, p.xvi) summary of Gul and Pesendorfer's captures this nicely:

> "[According to Gul and Pesendorfer,] [t]he traditional normative economic question concerning how to design policies to forward some definition of the good inappropriately treats economics as social therapy. The only significant role of normative economics should be to highlight interesting questions for positive economics, such as how best to model the forces accounting for apparent failures of Pareto optimality. Given that normative economics is designed only to suggest new approaches to positive economics, it too must be centered around theories of choice."

(e.g., the conventions of mathematics and logic). By taking into account arguments against this other dichotomy, it can be argued that the entanglement thesis is stronger than Mongin's weak non-neutrality thesis, yet weaker than the strong non-neutrality thesis: it is a *moderate* non-neutrality thesis. I do not suggest that Mongin is *necessarily* in disagreement with the entanglement thesis. Indeed, though he does not put them together with his arguments against the fact/value dichotomy (2006c), Mongin (2006b) has also advanced a set of arguments against the fact/convention dichotomy in economics (by bringing together arguments from the philosophy of language, of science and of economics, but not of mathematics and of logic). By contrast, as said in the general introduction the specificity of the entanglement thesis is to bring together arguments from various areas of philosophy against both the fact/value and fact/convention dichotomies to reshape the usual way of characterizing 'facts' by making explicit their interdependence, in science, with values and conventions.

Conclusion

This section is the only place in this dissertation where the positive/normative issue at the level of the articulation of positive and normative economics is discussed at some length. This discussion was necessary to fully understand the positive/normative issue at the level of models of individual behaviors on which the rest of the dissertation focuses. This is so because part of the normativity in the latter level comes from the former level. It was also necessary to fully appreciate a comparison between Sen and the behavioral economics of Thaler, since it is on the former level that one finds their strongest disagreements. The underlying methodological issues of these disagreements crystallize in the gap we found between behavioral economists' position and the entanglement thesis within the picture of the contemporary reflections on the matter. Despite our adherence to the entanglement thesis, it should be noted that Thaler's 'choice architects' in his contributions to normative economics allowed us to give further grounds to our decision modeler in the communicative structure of choices.

Conclusion and transition: catching the drifts towards the three dimensions

We have seen that Sen and Thaler (along with other behavioral economists) criticize the same standard models of individual behaviors in consumer choice theory and rational choice theory, with the same interpretation of the positive/normative issue within those models. Part of their methodology is similar, e.g., the uses of intuitive example, introspection and, most importantly, a nontrivial and explicitly stated uses of discussions with decision makers, i.e., what we called the methodological communicative structure of choices. We pointed the non-triviality of the latter by examining its empirical counterpart and some developments through the theory of speech acts. The key distinction was between, on the one hand, a decision modeler who poses a decision problem to, on the other hand, the decision makers traditionally studied in economics. The decision modeler is, on our account, very similar to Thaler's choice architects (i.e., can be identified with a firm, a policy maker etc.), albeit slightly more general as one motivation underlying its development was to capture interactions between economic agents and economists doing positive economics. We also saw how Sen and the psychologists who inspired the behavioral economics of Thaler, i.e., Kahneman and Tversky, proposed a set of non-mutually exclusive normative criteria to evaluate the content of preferences and how this relates to all these authors' criticisms of the methodological convention that preferences are exogenously 'given' in standard economic analysis.

However, despite these possible mutual developments Sen and the behavioral economics of Thaler share non-trivial disagreements. The basic one underlies their different methodological position regarding how the positive/normative issue in models of individual behaviors should be dealt with. Roughly, behavioral economics wishes to keep the standard model for its normative dimension but to strip it down of its positive dimension. Behavioral economists' models of individual behaviors inspired from, or directly using, *P*sychology, propose to take the lead on that latter dimension. On the other hand, Sen wishes to re-articulate the positive and normative dimensions within models of individual behaviors instead of separating these dimensions. The entanglement thesis plays a crucial epistemological justification here since Sen's methodology implies some violations of the fact/value dichotomy. Finally, we have shown that the implications

of these different methodological positions are at their strongest on the issue of the articulation between positive and normative economics.

Though the focus was on standard models of individual behaviors *under certainty*, at several points the discussion drifted towards three other dimensions of individual behaviors. This was notably the case while discussing arguments against the exogeneity of 'given' preferences. There, we saw that *the information* about the contents of preferences that could give some normative justification to their endogeneity were located in the relation of the decision makers to the dimensions of (1) other people, (2) time, (3) *un*certainty. The work of Sen strongly emphasized (1) to the point of seeming to be primary over (2) and (3), and roughly the same can be said about Kahneman for (2) and Tversky for (3). Clearly, the rather informal and brief nature of these drifts of the discussion towards these three dimensions calls for more careful scrutiny. This is what I propose to do in the next chapter.

Chapter 2

Under uncertainty, over time and regarding other people: rationality in 3D

> "[A] rich toolbox for "neoclassical repair," [...] may be a curse (a Pandora's box) rather than a blessing [...]. Some guidance on how risk attitudes, time preferences, and other-regarding concerns are interrelated becomes, therefore, necessary when we want to make sound behavioral predictions. [...] Except for a few attempts, economic theory offers no idea of whether risk aversion goes hand in hand with patience and other-regarding concerns." (Güth et al. 2008, p.261 fn1 omitted)

Behavioral economics is often presented as having demonstrated inconsistencies of individual behaviors (see, e.g., Camerer and Loewenstein 2004). Besides the contributions under certainty discussed in the previous chapter, if one asks 'but inconsistency with respect to what exactly?', answers are likely to be at least threefold, e.g., "risk attitudes, time preferences, and other-regarding concerns" in the words of Werner Güth and his co-authors. Depending on the tar-

geted audience, these are expressed differently. Using much broader terms to reach an audience outside of economics, Thaler and his co-authors have stressed "Three Bounds of Human Nature" (Mullainathan and Thaler 2001, p.1095), namely that "people exhibit bounded rationality, bounded self-interest, and bounded willpower" (Jolls, Sunstein and Thaler 1998, p.1471). By contrast, addressing economists explicitly, Stefano DellaVigna announces the first section of his influential survey of behavioral economics using the language of preferences : "[t]he first class of deviations from the standard model [...] is nonstandard preferences [...] on three dimensions: time preferences, risk preferences, and social preferences" (2009, p.316). In short, behavioral economics has mainly challenged standard models of individual behaviors under uncertainty, over time and regarding other people.

This chapter has four related goals – the first three are derived from the general perspective of this dissertation and the last one is more specific to this chapter. The first goal, pertaining to the positive/normative issue, is to make explicit the entanglement of facts, values and theoretical conventions constitutive of the main challenges posed by behavioral economics to the standard accounts of behaviors under uncertainty, over time and regarding other people. This will be done in roughly the same fashion as in the previous chapter, i.e., by scrutinizing empirical and theoretical contributions through the entanglement thesis (though with a less systematic comparison to Sen's work here). The second goal, pertaining to the issue of interdisciplinarity, is to study the relations between economics and *p*-&-*P*sychology underlying these challenges. The third goal, pertaining to the role of language in economic rationality, is to make explicit the communicative structure of economic choices that underlies these challenges. To do so, I will draw on Table 2.1 taken from linguist-semiotician François Rastier, which organizes some similarities of ordinary language uses *across* the three dimensions by three zones *within* them.[1]

Rastier capitalizes the words in the table because they represent specific types of information that can be instantiated in communication through a variety of other words, e.g., 'SURE', 'LIKELY', 'IMMEDIATELY', 'LATER', 'ME', 'US', 'DEAR READER', etc. Throughout this chapter, we shall stick to Rastier's typographical convention of capitalizing the information about the

[1] This is a modified version of Rastier's (1996, fig. 14.2; see also 2008, p.213; 2012, fn41 p.23). There are two main modifications. First, what Rastier calls "*Mode*" what is here called '(Un)certainty', and with less difference in translation, "*Personnes*" what is here called 'People'. Second, a fourth dimension in his table, 'Space' (with HERE, THERE and OVER THERE/ELSEWHERE as its three zones), is not a relevant dimension of individual behavior in this chapter, and has thus been omitted

	Identity Zone	Proximal Zone	Distal Zone
(Un)certainty	Certain	Probable	Possible, Unreal
Time	Now	Once, Soon	Past, Future
People	I, We	You	He, It

Table 2.1: Similarities of language uses across the three dimensions

three dimensions, especially to make such information explicit when they are implicit. This will be done to illustrate the following claim: one condition of possibility for the challenges posed by behavioral economics on the three dimensions is that a decision modeler can use ordinary language to mark the distinctions of this table, so as to communicate them to the decision maker. In other words, marking the distinctions within and across Rastier's table is a condition of possibility for most if not all of the issues discussed in this chapter to be intelligible in the first place. Showing this throughout this chapter is the main way by which the role of language in economic rationality will be tackled. Finally, the last (and more specific) goal of this chapter is to provide systematic discussions of the three dimensions *altogether* – to see, as it were, rationality in 3D.[2]

By systematically seeing rationality in 3D, the intended contribution of this chapter is to provide a better understanding of the issues underlying the contemporary behavioral *versus* standard economics debates. The evolution of the literature around behavioral economics on the three dimensions followed a pattern that can be nicely interpreted through Rastier's table. Prior to the mid-2000s, interactions between zones of a given dimension were studied, e.g., between the identity and distal zones *within* the dimension of uncertainty in pairs of decision problems such as (Kahneman and Tversky 1979, p.266):

€3000 For sure *versus* 80% chance of getting €4000
25% chance of getting €3000 *versus* 20% chance of getting €4000.

The research agenda focused on how such interactions create anomalies for standard models of risk, time and social preferences, *respectively*, and on developing alternative 'behavioral' models to offer parsimonious accounts of these dimension-dependent anomalies. From the mid-

[2] Chen (2013) provides a comparative study of saving behaviors across countries that have different language in the sense that different zones in the dimension of time are or are not linguistically marked by verb conjugation. He finds that around 30% of the differences between saving behaviors are statistically explainable by these linguistic differences.

2000s onwards, interactions across dimensions started to be studied, e.g., between the identity and distal zones *across* the dimensions of uncertainty and time in pairs of decision problems such as (Baucells and Heukamp 2010, p.151):

€9 For sure In three months *versus* 80% Chance of getting €12 In three months
10% Chance of getting €10 In three months *versus* 8% Chance of getting €12 In three months

As the epigraph from Güth et al. (2008) suggests, the topic of interactions across dimensions has been understudied in economic theory. Even the contributions on the interactions across dimensions are quite disconnected from one another. That is, each of these contributions focuses on *one* of the three pairs (time-risk *or* risk-other people *or* other people-time) and usually in *one* direction (e.g., risk introduced in intertemporal problems *or* time introduced in problems under uncertainty *or* risk introduced in interpersonal problem etc.). Not only the three dimensions are never tackled altogether, but different pairs of cross-interactions are never contrasted with one another. It can be argued that the widespread urge to *isolate* the three dimensions from one another turns this tripartition into a trichotomy, which is as theoretically counterproductive as taking the fact/value distinction as a dichotomy was argued to be methodologically counterproductive in the previous chapter. Trying to systematically see rationality in 3D is intended to counterbalance this widespread urge for isolation.

More precisely, the intended contributions with respect to the contemporary behavioral *versus* standard economics debates are the following. As we shall see, while interactions within dimensions created anomalies for standard models but were in line with behavioral models[3], some introductions of extra dimensions created anomalies for behavioral models but were in line with standard models[4]. Seeing rationality in 3D points a non-trivial need for convergence between standard and behavioral economics' models if all the anomalies are to be parsimoniously captured theoretically. Furthermore, a convergence is more likely to occur with the three types of preference taken as theoretical targets, rather than other *P*sychological constructs such as 'emotion', 'attention' or 'motivation'. This is so because risk, time and social preferences have a

[3]e.g., [€3000 For sure $\succ$ 80% chance of getting €4000]&[25% chance of getting €3000 $\prec$ 20% chance of getting €4000]

[4]e.g., [€9 For sure In three months $\prec$ 80% chance of getting €12 In three months]&[10% chance of getting €10 In three months $\prec$ 8% chance of getting €12 In three months]

long history in, and are central to, standard economics in a way other *P*sychological constructs are not (which does not imply that the latter cannot play any role in economic theory).

There are two limits of this chapter that are worth noting before the conclusion. The first limit is that this chapter focuses much more on what could be called the *qualitative* challenges from behavioral economics than on the *quantitative* ones. By qualitative challenge I mean inconsistent behaviors that violate some axioms underlying the standard models, making it impossible to represent those behaviors by *one* utility function. Quantitative challenges, on the other hand, are behaviors that can be captured by one such function, but at the price of implying absurd numerical values and degrees of curvature (interpreted as *measures* of risk, time *and/or* social preferences in different situations, see Broome 1991). I focus mainly on the qualitative issues because it can be argued they furnish the conditions of possibility for the quantitative ones. Of course both are related and play a role in the standard *versus* behavioral economics debates. I will only discuss the quantitative ones that arguably play a non-trivial role in the debates on which this dissertation focuses, and briefly so, i.e., without fully characterizing their relations with the qualitative ones. The second limit is that the traditional distinction in economics between probabilistic 'risk' (the proximal zone of uncertainty in the table) and non-probabilistic 'uncertainty' (its distal zone in the table) will not be terminologically respected in this chapter, as the latter will not be treated so that both terms will be used synonymously to refer to the former. Thus the distal zone in the dimension of uncertainty will not be discussed, i.e., decision making regarding what Possibly happens (though not probabilistically so, see the literature on ambiguity, e.g., Wakker 2010), or what may be thought of as Unreal (see the literature on ignorance and unawareness, e.g., Zeckhauser 2014 provides some references and claims that this is where the future of the economics of uncertainty lies to understand the real world).

Finally, one precision needs to be made. Though the three dimensions are always discussed altogether in all the subsections, the latter are systematically structured by first discussing risk preferences, then time preferences, and finally social preferences. This is done for two main reasons, reflecting two related senses of what could be called a primacy of risk over time over social preferences. Firstly, risk preferences have been central in the making of behavioral economics, especially regarding its treatment of the normative/positive distinction in models of individual

behaviors (see Heukelom 2014). Even if this is now also the case for time preferences, historically, discussions of risk preferences influenced discussions of time preferences more than the reverse in the making of behavioral economics. And discussions of social preferences have always been less connected to either one of the two others than the two others have been connected between themselves. Secondly, in standard economics (and outside of decision theory) risk preferences seem to be far more used and taken to be far less problematic than time preferences (see, e.g., Gollier 2001), while social preferences seem to be always controversial (see, e.g., Mongin and d'Aspremont 1998; or Binmore and Shaked 2010a;b *versus* Fehr and Schmidt 2010 and Eckel and Gintis 2010). We shall see that this implicit primacy of risk over time over social preferences is also reflected in the pairs of dimensions that have attracted the most attention in contributions on interactions across dimensions.

This chapter is simply structured in two sections that mimic the before/after mid-2000s movement mentioned above, i.e., first discussing behavioral economics' challenges stemming from interactions within dimensions (2.1) and then from interactions across dimensions (2.2).

2.1 Interactions within the three dimensions

A picture of the main empirical regularities that have motivated behavioral economists' work on the three dimensions is first presented (2.1.1), and shown to violate instantiations of 'consequentialism' in the standard models of preferences for the three dimensions (2.1.2). The three primacies from the previous chapter are then developed altogether to suggest how they may offer counter-arguments to the standard consequentialist picture of the three dimensions (2.1.3). Finally, the main theoretical developments that the main empirical regularities have motivated in behavioral economics are discussed altogether (2.1.4), under the critical scrutiny of the arguments from the previous subsections.

A general note on the way experiments will be discussed throughout this dissertation is in order. Since most experiments involve monetary consequences, all currencies are displayed in euros and some slight rewordings and changes in the letters that denote objects of choice are also made. All this is done to homogenize the decision problems and facilitate the reasoning regarding the theoretical implications. Furthermore, while most, if not all, experiments presented here are

considered as robust in behavioral economics, we are not concerned with issues of replications and statistical significance (e.g., whether 51% or 100% of the subjects of an experiment displayed the effect): the focus is on what these behavioral patterns can tell us, *theoretically speaking*, about the standard *versus* behavioral economics debates.

2.1.1 The main empirical regularities, altogether and naked

This subsection tries to present a structured picture of the main empirical regularities that are being discussed around behavioral economics with respect to the three dimensions. Compared to what is usually presented in the literature (e.g., Prelec and Loewenstein 1991), the two main specificity of this picture is that (1) it focuses on the similarities and differences of the three dimensions *altogether*, and it is (2) *naked* in the sense of being devoid of theoretical and methodological considerations. The former specificity is self-justified from the goal of this chapter, while the latter, which looks quite at odds with the entanglement thesis, is justified from both convenience (allowing easy references in later subsections to the 'facts' under discussion) and conciseness (not disrupting the discussion with theoretical and empirical comments increases the salience of the similarities and differences between dimensions). It should also be noted that some of the examples illustrating the regularities have been carefully chosen for the way they make the 'logic' of the regularities transparent, though they are extreme instances of them.

The example of a decision problem under risk mentioned in the introduction was part of what is usually labeled the '*common ratio effect*' (more precisely, a specific version of it called the '*certainty effect*') from Kahneman and Tversky (1979; problems 3 and 4). The decision maker is presented the following two pairs of objects of choice (*A versus B*, and *C versus D*). The first one was mentioned in the introduction, it is between objects of choice with a CERTAIN consequence (3000€) or PROBABLE consequences (4000€ and 0€):

> **A**: The certainty of winning €3000 [i.e., 100% chance]
> *B*: 80% chance of winning €4000 [and 20% chance of winning nothing]

The second pair is obtained from the previous one by reducing the probability of occurrences of the non-null consequences by a common factor, here by 4, so as to choose between objects of choice with only PROBABLE consequences:

C: 25% chance of winning €3000 [and 75% chance of winning nothing]
D: 20% chance of winning €4000 [and 80% chance of winning nothing]

For the time dimension, consider the following '*common difference effect*' (more precisely, a specific version of it called the '*immediacy effect*') from Thaler (1981, opening example). Again there are two pairs of objects of choice. The first pair is between objects of choice with consequences occurring NOW or LATER:

A: One apple today [and nothing tomorrow]
B: Two apples tomorrow [and nothing today]

The second pair is obtained by keeping the consequences, but increasing the delay of the non-null consequences by a year, so that they are occurring in the FUTURE though one even LATER than the other:

C: One apple in one year
D: Two apples in one year plus one day

For the social dimension, consider the following '*dictator game*' from Kahneman, Knetsch and Thaler (1986b, experiment 2, part 1). The decision maker is given a choice between two ways of dividing €20 between HIMSELF and ANOTHER anonymous person:

A: €18 to self and €2 to other
B: €10 to self and €10 to other

Before we present more examples, a notational and terminological note is helpful at this point to highlight some similarities and differences between the three dimensions. An object of choice X can be represented as a set of consequences $\{x_1, x_2, ..., x_n\}$ *distributed* either probabilistically ($p_1, p_2, ..., p_n$ summing to 1), or over time $(t_1, t_2, ...)$, or across people $(i, j, k, ...)$. In the standard terminology, objects of choice can be 'lotteries' where x_1 has p_1 chance to occur, etc. so that $X = (x_1, p_1; ...; x_n, p_n)$; or 'plans' where x_1 happens at t_1, etc. so that $X = (x_1, t_1; x_2, t_2; ...)$; or 'allocations' where x_1 is for individual i, etc. so that $X = (x_1, i; x_2, j; ...)$. For instance, $B = (b_1, p_1; b_2, p_2) = (€4000, .8; €0, .2)$ in the common ratio

effect, $B = (b_1, t_1; b_2, t_2) = (0\,apple, today; 2\,apples, tomorrow)$ in the common difference effect, and $B = (b_1, i; b_2, j) = (€10, self; €10, other)$ in the dictator game.

Following Drazen Prelec and Loewenstein (1991), who already inspired the presentation of the first two examples on the dimensions of uncertainty and time, a structured way to introduce more examples is to present what they call '*sign effects*' and '*magnitude effects*' (1991, pp. 774-5). An instance of a sign effect under uncertainty that is especially convenient to introduce here is the following one, taken from Kahneman and Tversky (1979, problems 3' and 4'). Take the above example of the certainty effect, and simply reverse the sign of the consequences from monetary gains (e.g., winning €3000) to monetary losses (e.g., losing €3000):

A': The certainty of losing €3000 [i.e., 100% chance]
B$'$: 80% chance of losing €4000 [and 20% chance of losing nothing]
C$'$: 25% chance of losing €3000 [and 75% chance of losing nothing]
D': 20% chance of losing €4000 [and 80% chance of losing nothing]

Notice that here the pattern of preferences $[B' \succsim A']\&[C' \succsim D']$ is the reverse of the one over monetary gains $[A \succsim B]\&[D \succsim C]$; Kahneman and Tversky (1979, p.268) label this sign effect under uncertainty the 'reflection effect'. Though Thaler (1981) demonstrates some sign effects over time, the following other instance is easier to introduce (it is taken from Faralla, Benuzzi, Nichelli and Dimitri 2012, stimuli 107 and 109). In a first set of pairs of plans (*A versus B* and *C versus D*), the same pattern of preference as in Thaler's apple example is observed, i.e., a preference for a small consequence NOW over a slightly bigger one LATER, which reverses when the first consequence is available SOON and the second one even LATER:[5]

A: Winning €5 today
B: Winning €7.50 in a month
C: Winning €5 in two weeks
D: Winning €7.50 in a month and two weeks

Then, by simply reversing the sign of the consequences, from monetary gains to monetary losses, one can observe:

[5]Faralla et al. (2012) have very clean experimental tasks, making it easy to introduce the patterns of preferences constitutive of the effects under discussion. But their discussion of the data is not at the level of comparisons between pairs of plans, as the bolding of modal preferences here would suggest. It is however in line with the authors' qualitative discussion of the patterns in their data, and of course in line with the sign and magnitude effects as defined by Prelec and Loewenstein.

A': Losing €5 today
$\mathbf{B}'$: Losing €7.50 in a month
$\mathbf{C}'$: Losing €5 in two weeks
D': Losing €7.50 in a month and two weeks

With this pattern of preference just displayed, it is easy to introduce an instance of magnitude effects over time. The following pattern of preferences are between pairs of plans which have the same delay as in the previous sign effect example, but the amount of money (i.e., the *magnitude* of the consequence) is increased by a factor of six (Faralla et al. 2012, stimuli 117 and 119):

A'': Winning €30 today
$\mathbf{B}''$: Winning €45 in a month
C'': Winning €30 in two weeks
$\mathbf{D}''$: Winning €45 in a month and two weeks
$\mathbf{A}'''$: Losing €30 today
B''': Losing €45 in a month
$\mathbf{C}'''$: Losing €30 in two weeks
D''': Losing €45 in a month and two weeks

Therefore, the magnitude effect over time is the overall pattern of preference whereby common difference effects tend to disappear when the magnitudes of the consequence increase, for both gains and losses.

As an instance of a magnitude effect under uncertainty, Prelec and Loewenstein (1991) defer to a series of informal decision problems that Harry Markowitz (1952, pp.153-2) posed to his friends. Markowitz reports to have observed the following pattern of preferences, whereby the Probability of getting a very small consequence is preferred to the Certainty of receiving an even smaller (by a factor of 10) consequences, but the Certainty of receiving a very big consequence is preferred to the (same) Probability of receiving an even bigger (still by a factor of 10) consequences:

A: The certainty of winning 10 cents
$\mathbf{B}$: 10% chance of winning €1
$\mathbf{C}$: The certainty of winning €1,000,000
D: 10% chance of winning €10,000,000

And such a magnitude effect is also observed for monetary losses, though the pattern of preferences is the reverse of the one for gains:

$\mathbf{A}'$: The certainty of losing 10 cents
B': 10% chance of losing €1
C': The certainty of losing 1,000,000
$\mathbf{D}'$: 10% chance of losing €10,000,000

This concludes the presentation of empirical regularities on two (uncertainty and time) of the three dimensions for this subsection. What are the corresponding effects in the third dimension, that of other people? It is possible to make explicit parallels between the structure of the effects on risk and time preferences just discussed and some of the classical challenges from the literature around behavioral economics that stemmed from experiments on the dictator game. By contrast with the original dictator game from Kahneman et al. (1986b) presented above, most experiments do not present a decision problem between two allocations. Rather decision makers can allocate a given amount of money between HIMSELF and THE OTHER as he wishes (i.e., he can choose within a quasi-continuous distribution of monetary consequences).[6]

The simplest parallel suggesting itself is with magnitude effects. By contrast with the previous examples, changing the amount of money to be allocated in a dictator game does not usually have an effect on the chosen allocation. Notice that it is *the absence* of an effect that is taken as a (problematic) empirical regularity (e.g., List and Levitt 2007, p.164; and see the references in Novakova and Flegr 2013). In one of the rare study displaying a magnitude effect, Julie Novakova and Jaroslav Flegr (2013, Table 1) strongly increased the amount of (hypothetical) money to be allocated: four times by a factor of 10 – e.g., €20, €200, €2000, €20000, €200000; the proportion of the allocation given to THE OTHER reduced constantly from 28.3% to 23.3%. Even in that case, the allocations to THE OTHER are well within the 20-30% range of aggregate results on standard dictator game, with less than 40% of decision maker giving nothing (see Camerer 2003, chap.2; Engel 2011). In other words, even very large magnitudes of monetary consequences do not drive decision makers to offer nothing or close to nothing.

Though the parallel with money is easy because it is identical in experiments on the three dimensions, the parallel between uncertainty or time, on the one hand, and other people, on the

[6]Other games with more game-theoretic issues (e.g., ultimatum games, trust games etc.) that are used to study social preferences, but they will not be discussed in this dissertation. We focus on the dictator game because the degree of interactions between the players are very weak. Game-theoretic issues render the comparison with the two other dimensions less clear because the latter underlie more decision-theoretic issues (though see Aumann and Dreze 2009), and they underlie interactions of risk and time preferences which would complete the comparison but again at the cost of some complications (see the role of 'strategic uncertainty' in 'repeated games' in Camerer 2003).

other, is less straightforward. What has been called '*social distance*' in the literature around the dictator game provides a plausible parallel with uncertainty or time. A traditional dictator game always involves the three zones in the dimension of people: a decision maker (I, The allocator) communicating to the experimenter (You, The economist) how much to give to someone else (He, The recipient) who is usually anonymous. In standard designs, there is close to zero communication between the allocator and the recipient; though both communicate with the experimenter. Roughly, the social distance in a dictator game characterize the communication among these three protagonist, whether and how they talk to each other and what they know about each others' identity and actions. The challenging empirical regularities in this literature is that variations of social distance can push allocators *both ways*: either to give more than the 20-30%, or, at the opposite, to keep more than 80-70% (see Engel 2011, sect. 4.5). Two classical papers illustrate this point. Notably using double blind experimental design where both the experimenter and the recipient do not know who gave to whom, Elizabeth Hoffman, Kevin McCabe and Vernon Smith (1996) show how the less communication between the allocator and the experimenter, the more the former is likely to keep 100% of the money for himself. Keeping communication between the allocator and the experimenter very low, Iris Bohnet and Bruno Frey (1999) show how the more communication between the allocator and the recipient (from the former just seeing the latter, to both seeing each others, through the former seeing and hearing personal information about the latter), the more the former is likely to propose above the traditional 20-30% of the allocation to the latter. *The identity of the recipient* is indeed one of the rare variable that seems to trigger (mean) allocation where the allocator keeps less than 50% for himself, as when the recipient is from a poor community in a third world country (Brañas-Garza 2006).[7]

As for the impact of monetary losses, it has not been as thoroughly studied in the dictator games as it has been in decision under risk and over time. The empirical regularity that can be interpreted as being about losses and that has been discussed around behavioral economics, is the following. Endowing both the allocator and the recipient with some money and allowing

[7] 'Social distance' was also used long ago by Edgeworth to discuss what is today called social preferences (see Fontaine 2000, p.413), and is currently used by behavioral psychologists making the case that it is the interpersonal analogue of probability and time in decision making (see e.g., Jones and Rachlin 2006; and Ida and Ogawa 2013 for a discussion in economics).

the former to give some of his money as well as to take from some of the latter's tend to push the former to give 0%, but not to take from the latter (List 2007; Bardsley 2008). But that does not (nor was it intended to) parallel the studies of losses under uncertainty or over time where some decision problems have no positive monetary consequences at all. Therefore potential sign effects in the dictator game are yet to be investigated (though see Antinyan 2014 in economics; and Leliveld et al. 2009, exp.3, in social psychology).

All the effects discussed in this section were meant to study the interactions of different zones within a given dimension in Rastier's sense. The implicit assumptions behind the experiments discussed here is that the Sure or Probabilistic consequences of a lottery are Immediate and for Oneself (i.e., not Delayed and/or for Someone Else); the Immediate or Delayed consequences of plan are Sure and for Oneself; and the consequences of an allocation between Oneself and Another are Sure and Immediate. The experimental literature on the dimension of other people points another implicit assumption, namely that the decision maker who reasons in the first person (I, Me) communicates in the second person (You, Mr. The Experimenter, or what-not) to a decision modeler who posed the decision problem under investigation (as it was emphasized throughout the previous chapter).

2.1.2 Violations of consequentialism

In discussions around behavioral economics, the patterns of choice just presented are usually taken as factual observations that are problematic regarding different theoretical conventions of the standard models of risk, time and social preferences, from which value judgments of rationality or irrationality are often derived. More precisely, these patterns of choice are problematic for the frameworks of expected utility (EU), exponentially discounted utility (EDU), and self-interest into which risk, time and social preferences can be respectively modeled. There is a striking contrast between the first two and the last one. On the one hand, the first two frameworks have been expressed in formal language through different uses of the axiomatic method at different times; in economics, the first ones being usually credited to Paul Samuelson (1937) or Tjalling Koopmans (1960) for EDU, and John von Neumann and Oskar Morgenstern (1947, chap.1; 1953, appendix) for EU. On the other hand, "the assumption of self-interest is not presented as a formal axiom – it does not come in, as it were, by the front door" (Walsh 1996,

p.113). As we shall see, systematic discussions of self-interest are nevertheless possible through Sen's (2002, chap.1, sects 8-10) critical characterizations of various ways by which self-interest enters standard models in economic theory by the back door. The goal of this subsection is to contrast the three main theoretical conventions violated by the patterns of preferences discussed in the previous subsection, namely, the *independence axiom* constitutive of EU, the *dynamic consistency* implied by EDU, and the *self-centeredness* underlying self-interest. The main point of connection emerging from these contrasts is what is usually referred to as the *consequentialism* of standard economics.

The theoretical framework of EU provides the conditions under which a decision maker chooses a lottery over another one, i.e., if and only if he prefers the chosen one and if and only if he derives more utility from the chosen and preferred one; or in formal language and with the last chapter's notation, the conditions under which $X = C(\{X, Y\}) \Leftrightarrow X \succsim Y \Leftrightarrow U(X) \geq U(Y)$. There are additional theoretical conventions from the framework of consumer choice theory due to the features of the objects of choices, i.e., of the probabilities in the lotteries. The expected utility ($U(.)$) of a lottery is a real number given by the sum of the evaluations of its consequences by a function ($u(.)$), linearly weighted by their respective probabilities. Expressed more formally, the expected utility of B in the common ratio effect above is calculated by $U(B) = u(b_1)p_1 + u(b_2)p_2 = .8u(€4000) + .2u(€0)$. The linear weighting by probability is represented by the absence of a function that could change the numerical values of the probabilities. The so-called 'meaning of $u(.)$', and especially its relation with $U(.)$ under certainty, is a matter of controversy that we will not discuss in this dissertation. And to avoid making controversial statements, we will call $u(.)$ a "*subutility function*" (following Gorman 1968, p.368), and say accordingly that it 'subevaluates' the consequences or give real numbers representing the 'subutilities' of the consequences, while $U(.)$ evaluates lotteries.[8]

The independence axiom, "the most critical condition in [the] behavioral foundations of expected utility" (Wakker 2010, p.36), is a condition that the decision maker's preferences over

[8]Roughly, the issue is that $u(.)$ has cardinal properties for comparisons among consequences that $U(.)$ does not need for comparisons among lotteries (or among temporal plans or social allocations for the two other dimensions, see below). So if both are strongly related, then we may end up with cardinal utility in decision under certainty, *contra* the ordinalist revolution (see esp. Fishburn 1989b; Wakker 1994; Ellingsen 1994; Guala 2000; Abdellaoui et al. 2007; Mongin 2009; Moscati 2013a; b; 2016a; b). This gets even more problematic if the cardinality of $u(.)$ is given explicit *p*-or-*P*sychological meanings (contra purely techincal ones), as is the case in behavioral economics (see Rabin 2000, fn3 p.1282; Wakker 2010).

lotteries have to respect for the key theoretical convention $A \succsim B \Leftrightarrow U(A) \geq U(B)$ to hold. The most usual way of stating the independence axiom is in terms of the independence of preferences from the 'mixing' of lotteries. $pA + (1-p)X$ is said to be a mix of lottery A with lottery X, and the mix can be weighted by any probability p. The independence axiom states that a preference for a lottery over another should not change if a third lottery is mixed with both former lotteries in the same way. For instance, if there is a preference between two lotteries $A \succsim B$, it should not change (for the mix where there is B over the mix where there is A) if both are mixed with X with the same weight: $pA + (1-p)X \succsim pB + (1-p)X$. Thus the independence axiom can be expressed as $A \succsim B \Leftrightarrow pA + (1-p)X \succsim pB + (1-p)X$. In the common ratio example, we can denote by X the certainty of getting nothing, i.e., (with abuse of ordinary language) the 'lottery' $(x_1, p_1) = (0, 1)$, and mix it by the same weights of $p = .25$, $1 - p = .75$ with $A = (€3000, 1)$ and $B = (€4000, .8; €0, .2)$ to get $.25A + .75X = C = (€3000, .25; €0, .75)$ and $.25B + .75X = D = (€4000, .2; €0, .8)$. Thus the modal pattern of preference in the common ratio effect $[A \succ B] \& [D \succ C]$ violates the independence axiom; magnitude and sign effects under uncertainty also violate the independence axiom.[9]

The theoretical framework of EDU can be stated in quite similar terms. The discounted utility ($U(.)$) of a plan is a real number given by a weighted sum of the subutilities of the consequences. However, unlike the linear weighting by probabilities in EU, the weighting by time in EDU is exponential through a discount function ($\delta(.)$). From the period $t-1$, the subutilities of the consequences that happen at t are discounted by $\delta^{t-1} = (\frac{1}{1+\rho})^t$, where ρ is the *constant* rate of discounting that the decision maker uses indifferently in all periods. For instance, in the apple example from Thaler (1981), $U(B) = \delta^{1-1}u(b_1) + \delta^{2-1}u(b_2) = u(0\,apple) + \frac{u(2\,apples)}{1+\rho}$. Note that δ^0 always equals 1, so that the subutility of the first consequence is never discounted, and that of the second one is discounted by $1+\rho$, that of a third one would have been discounted by $(1+\rho)^2$ (though the discounting between the second and third periods is still $1+\rho$), that of a fourth one by $(1+\rho)^3$ (thought the discounting between the third and fourth periods is still $1+\rho$), and so on exponentially. ρ is often called 'pure time preference' and said to represent the decision maker's *patience* (when it tends to 0 or becomes negative) or impatience (when

[9] This is the least general way of presenting the independence axiom. Indeed, there are several ways by which the independence axiom can be stated formally and informally (see MacCrimmon and Larson 1979; Fishburn and Wakker 1995; and Mongin 2009), or decomposed in sub-parts (see Machina 1989; and Burghart et al. 2015).

it grows). Notice the subjective attitude to time by contrast with the objective perception of probabilities in EU.

By contrast with the independence axiom in EU, dynamic consistency is *an implication of axioms* in EDU. This implication is the consistent evaluation of the same plan through time, in the sense that its discounted value does not depend on the period from which it is evaluated. In the apple example, to evaluate C and D *from* $t = 1$ consist in evaluating 'one apple *at* $t = 365$ (one year) but nothing at $t = 366$' and 'nothing at $t = 365$ but two apples at $t = 366$', respectively. On EDU's terms, the modal preference is $D \succ C \Leftrightarrow U(D) > U(C)$. To evaluate C and D *from* $t = 365$ consists in evaluating 'one apple today and nothing tomorrow' and 'two apples tomorrow and nothing today', respectively, i.e., it consists in evaluating A and B, respectively. On EDU's terms, the modal preference is $A \succ B \Leftrightarrow U(A) > U(B)$. Denoting the time of evaluation by a superscript (U^t), we have $U^{365}(C) = U^1(A)$ and $U^{365}(D) = U^1(B)$. Thus, time consistency requires that $U^1(C) = U^{365}(C) = U^1(A)$ and $U^1(D) = U^{365}(D) = U^1(B)$, which together imply $U^1(D) \geq U^1(C) \Leftrightarrow U^{365}(D) \geq U^{365}(C) \Leftrightarrow U^1(B) \geq U^1(A)$, which is obviously violated by the modal pattern of preferences in the common difference effect $[D \succ C]\&[A \succ B]$; magnitude and sign effects over time also violate dynamic consistency.[10]

Self-interest is not a theoretical framework in the sense EU and EDU are, mainly because the former is not an axiomatic system as the latter two can said to be. Compared with EU and EDU, issues around self-interest have been less around the formal language of optimization than around the ordinary language used to characterize the behaviors under study. This was pointed in the previous chapter when we discussed Sen's characterization of the three implicit requirement imposed by the notion self-interest in standard economics (self-centered welfare, self-welfare goal, self-goal choice). It is only one part of these requirement that is violated in dictator games, namely self-centered welfare, i.e., "[a] person's welfare depends only on her own *consumption* and other features of the richness of her life (without any sympathy or antipathy towards *others*, and without any *procedural concern*)" (Sen 2002, chap.1, p.33, my emphasises).

Violations of self-centered welfare in dictator games can be interpreted in a notation related

[10]More precisely, the two main axioms implying dynamic consistency are 'separability', which has the same mathematical structure as the independence axiom and gives EDU its separably additive form, and especially 'stationarity', which requires the subutility of a given consequence to be the same for any period; the latter is obviously central in yielding dynamic consistency (see Fishburn and Rubinstein 1982; Bleichrodt et al. 2008; Lapied and Renault 2012).

to the ones used so far, provided we make the standard identification of 'welfare' and 'utility' – thus we can call the requirement just *self-centeredness*. The self-centered utility of an allocation evaluated by ONESELF (i) is a real number given by a functional ($U^i(.)$) constituted by the sum of the subutilities of its consequences for ONESELF ($u_i(.)$) and OTHER PEOPLE ($u_{-i}(.)$, with $-i = j, k... \neq i$), with the requirement that for any consequence x we have $u_{-i}(x) = 0$, assuming otherwise the kind of standard conditions on preferences used in consumer choice theory (see Sobel 2005, sect.3.1). In Kahneman, Knetsch and Thaler's dictator game, self-centered utility implies $U^i(A) = u_i(a_1) + u_j(a_2) = u_i(€18) > U^i(B) = u_i(b_1) + u_j(b_2) = u_i(€10)$, which the modal pattern of preference $B \succ A$ violates; indeed any dictator game where the decision maker allocates even the tiniest amount of money to the other although he could have allocated it for himself violates self-centeredness. That violates only the "consumption" and "others" parts of the requirement of self-centeredness (italicized in the above quote from Sen). Pretty much all other variations in experimental design that trigger variations in allocations violate the "procedural concern" part of the requirement. It is not difficult to see that behaviors in dictator games do not necessarily violate self-welfare goal and self-goal choice, provided we drop the dubious but also standard identification of welfare and/or utility with choice, respectively.

With respect to the perspective of this chapter, the following quantitative issue about the standard accounts of the three dimensions should be briefly noted. Depending on the shared interests of economists working within *different* subfields, the curvature of $u(.)$ is interpreted as the marginal utility of consequences and allows inferences of *different* unobservable behavioral information. Among these behavioral information, three prominent ones are (1) risk aversion, (2) elasticity of intertemporal substitution and (3) inequality aversion. Informally, they respectively measure *the attitudes of decision makers towards more or less equal distributions of consequences* over (1) different probabilistic states of the world, (2) different points in time and (3) different people. In other words, there are different interpretations of the same mathematical construction. Most economists seem to go along with this by focusing, in a given situation, on one of the three dimensions while explicitly abstracting from, or assuming no specific role played by, one of the two others (see esp. Strotz 1956, p.166, focusing on time abstracting from risk or Pratt 1964, p.123, focusing on risk with no specific role for time); but one of the three is simply ignored (outside of normative economics and subfields such as the economics

of the family, of charity or of discrimination, the ignored dimension is usually other people). This strategy is worth noting for the purpose of this chapter only in order to provide some contrasts with two other strategies that aim to go beyond it. On the one hand, some economists think that there are good reasons to separate the dimensions and have accordingly specified more general functional forms to give different measures to two of the three different attitudes (see esp. Kreps and Porteus 1978 on risk and time preferences). On the other hand, other economists have searched normative justifications for why the three types of attitudes should be identical, and in the empirical cases they are not, which one among the three should play the normative benchmark defining how one of the other two should change for the decision maker to be totally consistent (see esp. Harsanyi 1955; 1988 on risk preferences as a benchmark for social preferences). Notice that in both cases one of three dimensions is ignored. These two strategies involve a mix of quantitative and qualitative considerations that hinges on the issue of interactions across dimensions; their discussion is therefore deferred to the next section.[11]

Only one of these considerations is appropriate to be discussed here, namely how the independence axiom, dynamic consistency and self-centeredness are all justified by value judgments derived from *consequentialism*, i.e., the principle that the consequences of a choice are the only sources of reasons from which that choice can be justified. It is rather immediate to see why this is so for these three requirements independently of each others. Self-centeredness is justified by consequentialism through the part of its requirement that "procedural concern" should not influence an individual's choices. The independence axiom and dynamic consistency are justified by consequentialism in the sense that they both force decision makers to reason only about the consequences of choices abstracted from *p*sychological inclinations and situational factors. Such implicit consequentialism is critically discussed by Sen, who proposes a more nuanced and less narrow version of consequentialism in various contributions (see the pages referenced at "consequences, relevance of" in the index of Sen's 2002 *Rationality and Freedom*). The role played by consequentialism in standard models of individual behaviors has been made formally

[11] For other instances of explicit discussions of interactions across dimensions, see the penultimate footnote in the previous subsection on 'strategic uncertainty' and 'repeated games'; Christian Gollier (2001) on the interactions between the dimensions of risk and time in macroeconomics and financial economics; the references at the end of the next subsection on the 'veil of ignorance' arguments in normative economics (that exploit expected utility to draw conclusions about social arrangements); Thibault Gajdos and John Weymark (2012) on the parallels between second-order stochastic dominance of distributions of income (hence related to inequality) and of probability; and the references in the last subsection of this section on intergenerational altruism in macroeconomics that have been borrowed by behavioral economics to model individual time preferences.

explicit over the years by Peter Hammond (1976; 1977; 1983; 1987; 1988a; b; 1989; 1998; Hammond and Zank 2014). Roughly, Hammond formally defined consequentialism as an axiomatic requirement on choice functions to show how this requirement *implies* dynamic consistency, the independence axiom and self-centeredness. What is worth noting about Hammond's contributions is that the notion of dynamic consistency becomes the normative benchmark from which risk and social preferences are to be evaluated, even in atemporal settings (i.e., when working on unreduced decision trees for decision under uncertainty and when comparing *ex ante* and *ex post* evaluations of social allocations in normative economics). Hence, while the primacy of risk over time over social preferences mentioned in the introduction of this chapter holds at a general level in the practice of standard economists, it is to be contrasted with a primacy of time over risk over social preferences that results from consequentialism as *the* normative justification of standard models of individual behaviors.[12]

Notice that in terms of Rastier's table, one condition of possibility for these discussions to have happened is the marking of distinctions *within* the three dimensions (and their subsequent articulations *across* them). In the words of the entanglement thesis, the *factual* patterns of choice that violate the theoretical *conventions* of the independence axiom, dynamic consistency and self-centeredness are subjected to *value* judgments from consequentialism. Put simply, criticisms of these patterns of choice as irrational are often made explicitly or implicitly from consequentialism, and defenses of these patterns of choice as rational or irrational often involve criticisms of (at least some features of) consequentialism. Sen's and Hammond's discussions of consequentialism point out several ways by which it is one of the deepest source of value judgments in economics. The origin of this is, according to them, mainly due to the philosophical influence of utilitarianism in economics (see also Walsh 1996). But is can be noticed further that, as logicians Dov Gabbay and John Woods (2005, p.21) put it: "[h]istorically, [...] logic is an examination of consequentialist reasoning whose success or failure is definable for or representable in semi-interpreted formal languages". Hence, besides the influence of utilitarianism, the widespread acceptance of consequentialist reasoning in standard economics can also be

[12]Peter Wakker pointed to me that, historically, the first contribution on what I call the primacy of time preferences, is the fifth chapter in the Burks, Arthur W. (1977) "Chance, Cause, Reason (An Inquiry into the Nature of Scientific Evidence)." The University of Chicago Press, Chicago. According to Wakker, "[t]his philosopher is not only the first, but also about the best and deepest on this topic" (personal communication 20/04/2016).

attributed to the epistemic value judgments favoring the uses of formal languages in economic theory, especially through the axiomatic method.[13]

Summing up, this subsection characterized the qualitative challenges, i.e., violations of the standard models' underlying axioms, posed by the empirical picture depicted in the previous section. The next subsection scrutinizes the main theoretical alternatives from behavioral economics to deal with these challenges.

2.1.3 Theoretical alternatives from behavioral economics

It is impossible to scrutinize the entanglement of facts, values and theoretical conventions underlying *all* these theoretical outcomes. This subsection concentrates on the most representative alternatives from behavioral economics, namely, *cumulative prospect theory* for risk preferences, *quasi-hyperbolic discounting* for time preferences and *non-self-centered motives* for social preferences.

Under risk[14]

Cumulative prospect theory (CPT) is the theoretical outcome of a series of historical episodes in decision theory where both the issue of interdisciplinarity between economics and psychology and the positive/normative issue loom large (see Heukelom 2014 for a more detailed historical account than the present one). In 1979, psychologists Kahneman and Tversky published their prospect theory (PT), which can be formally illustrated as follows. In PT, the utility of a lottery $X = (x_1, p_1; x_2, p_2)$ is a real number given by $U(X) = u(x_1)\pi(p_1) + u(x_2)\pi(p_2)$. In this formula, $u(.)$ is the value function presented in the previous chapter (i.e., concave for gains, convex for losses, and asymmetrically so around a reference point to represent loss aversion) and $\pi(.)$ is a *weighting function* that transforms *non*linearly the probabilities. Before Kahneman and Tversky (1979), economists displayed a *consequentialist sensitivity* through expected utility which transforms only the consequences but not the probabilities, and *vice-versa* for psychologists

[13] For a concise statement of the central issues underlying 'logical consequences', see Mikaël Cozic's (2009) presentation of a classical text from Alfred Tarski (2009 [1936]) and the text itself. Unrelatedly, it can be argued that economics sneaked in the rationality debates in *P*sychology and cognitive sciences (also related to Kahneman and Tversky's work), mainly through the door of consequentialism (see the economic examples and reasoning used by Baron 1994 in a debate within *P*sychology).

[14] Parts of the developments here have been used in Jullien (20016a).

who displayed a *probabilistic sensitivity* through the use of weighting function without nonlinear subutility functions (this contrast is borrowed from Wakker 2010, p.162). In a way, by $u(.)$ *and* $\pi(.)$, Kahneman and Tversky (1979) brought both sensibilities together. Economists have had issues bearing on the positive/normative distinction with both the weighting and value functions, which we briefly discuss in that order (see Wakker 2010 for detailed developments; see also Barberis 2013).[15]

Put simply, the weighting function transforms a probability p given by the decision modeler into the 'decision weights' $\pi(p)$ that the decision maker puts on the associated utility of consequences. Analytically, decision weights are not to be confused with subjective probabilities. One reason is that the former can be interpreted either as the irrational misperception of objective probabilities *or* as the rational over- or under-weighting of the importance of some consequences; subjective probabilities *may* be interpreted as irrational misperception (e.g., due to lack of information) but cannot be interpreted as over- or under-weighting because they have to sum up to 1. Exactly what shape the weighting function takes is an empirical and quantitative issue, where decision weights do not necessarily sum up to 1. To illustrate how this property account for some of the challenges presented above, consider the certainty effect, the probabilities of which are approximately plotted on Figure 2.1, i.e., on the graphical representation of Kahneman and Tversky's (1979) empirical estimation of the weighting function.

This 1979 version was conceived by Kahneman and Tversky *only for lotteries with at most two non-null consequences*. Furthermore, it is normalized at the end point $\pi(0) = 0$ and $\pi(1) = 1$, but not well-behaved near the end points 0 and 1, which can be interpreted as "the quantum of doubt" that makes it hard to draw exact lines representing "the categorical distinction[s]" between Certainty and Uncertainty, and between Uncertainty and Impossibility (Kahneman and Tversky 1979, p.282). Normalizing $u(€0) = 0$ to simplify the exposition, PT accounts for the certainty effect as follows. Though the decision modeler's presentation of the problems is 'the certainty of €3000 *versus* 80% chance of winning €4000' and '25% chance of winning €3000

[15] Peter Wakker pointed to me that he advocates, following Tversky, the use of 'prospect theory' for the 1992 version and of 'original prospect theory' for the 1979 version. According to them, the 'cumulative' adjective discourages people with no mathematical degree to use this version ("A theory with such a horrible technical name cannot carry far. let us use the best term for the best theory." personal communication 20/04/2016).

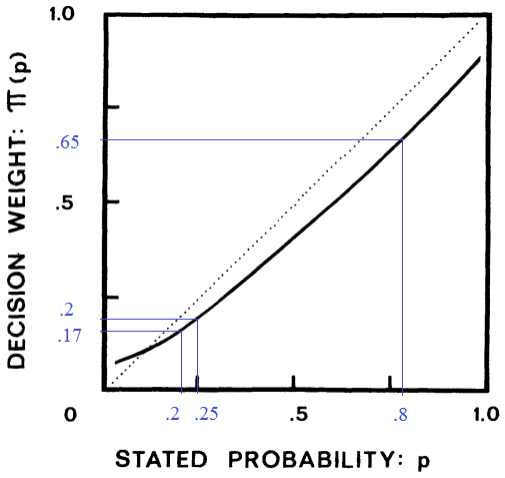

Figure 2.1: Kahneman and Tversky's 1979 weighting function and the certainty effect

versus 20% chance of winning €4000', the decision maker's representation of the objects of choice *when he evaluates* them is $[U(A) = u(€3000) > U(B) = .65u(€4000)]\&[U(D) = .17u(€4000) > U(C) = .2u(€3000)]$. Notice that if the decision weights were the given probability, there would be no common ratio effect as there is no common factor by which 1 is reduced to .2, and .65 to .17. Unlike in EU, the non-linearity of decision weights allows to formally represent the preferences that underlie common ratio (and other) effects. This was the main goal of Kahneman and Tversky, as it was for many other decision theorists.[16]

But the non-linearity of Kahneman and Tversky's (1979) decision weights also allowed to formally represent some preferences that standard decision theorists wish to avoid being formally representable, namely, preferences *against* first-order stochastic dominance. First-order stochastic dominance is maybe the most important theoretical convention in decision theory from which value judgments of rationality and irrationality are derived. It states that:

> "shifting probability mass from an outcome to a preferred outcome lead to preferred prospect. In other words, the more money the better." (Wakker 2010, p.65)

It thus applies only when there is a relation between two lotteries whereby there are more chance to win at least as much more money in the lottery that is said to first-order stochastically

[16]For reviews see Fishburn (1989a, sect.5), Machina (1987, pp.132-6; 1989, table 1), Camerer (1995, table 8.1), Starmer (2000, sect. 4.1), Quiggin (2014, sects. 12.4.2-12.7). A very pedagogical perspective is offered by Wilkinson and Klaes (2013, sect. 5.2).

dominate the other, e.g., '50% chance of winning €10 and 50% chance of winning nothing' first-order stochastically dominates '60% chance of winning €10 and 40% chance of winning nothing'. EU, as most of its alternatives in decision theory that are seriously considered by economists (cf. previous footnote), implies necessarily a preference *for* first-order stochastic dominance, i.e., for the lottery that first-order stochastically dominates. In other words, first-order stochastic dominance is "a hallmark of rationality" (ibid), and not necessarily implying a preference for it "has generally been regarded as a fatal flaw sufficient to damn theories of this type" (Bardsley et al. 2010, p130). Hence one reason why economists have had issues with PT is that despite its psychological advantages on the positive side, the non-linear decision weights do not necessarily imply preference for first-order stochastic dominance, hence it has some serious drawbacks on the normative side. This mainly comes from the non-additivity of decision weights, especially for small probabilities, e.g., $\pi(.01) + \pi(.06) > \pi(.07)$, so that '1% chance of winning €10, 6% chance of winning €10 and 97% chance of winning nothing' could be preferred to '7% chance of winning €11 and 97 chance of winning nothing' (for more detailed explanations see Kahneman and Tversky themselves 1979, pp.283-4; see also Wakker 2010, sect. 5.3).

The issue of interdisciplinarity between economics and *P*sychology looms large in how this issue was solved. Roughly, Kahneman and Tversky were in agreement with economists' stance on first-order stochastic dominance. Their way of bypassing the problem was to include in PT a set of rather informal considerations about the *p*-&-*P*sychology of decision makers to diminish the importance of the formal implications of violations of first-order stochastic dominance. Kahneman and Tversky (1979, p.274) avoid such violations by distinguishing "two phases in the choice process: an early phase of editing and a subsequent phase of evaluation". The latter *is* the value and weighting functions, that are therefore applied to the decision makers' representation of the objects of choices, which do not necessarily correspond with their presentation by the decision modeler (as in the above numerical example with the certainty effect). By contrast, the editing phase corresponds to how the decision maker transforms the decision modeler's presentation of the decision problem into a representation from which he will make his choice. Kahneman and Tversky only gave an informal account of it, in which two "operations" are worth mentioning for the purpose of this chapter (Kahneman and Tversky 1979, p.274). The first one is the "*Coding*" of consequences as gains or losses, i.e., the determination of a "neutral reference point" in the

value function, which is not necessarily 0 as we saw in the previous chapter (ibid). This aspect is discussed shortly. The second one is the "detection of dominance" which "involves the scanning of offered prospects to detect dominated alternatives, which are rejected without further evaluation" (p.275). This informal way of using the decision maker's *p*-&-*P*sychology to prevent a formal implication of a theory did not convince economists. What did convince at the same time standard economists, behavioral economists, and even Tversky and Kahneman (1992) is the so-called 'rank-dependent expected utility' or 'rank-dependent utility'[17] for short (RDU) approach pioneered by *economist* John Quiggin (1982). We shall not go into the formal details of this approach, but only note the methodological features of that matter for the purpose of this chapter.[18]

Quiggin formally modified psychologists' decision weights so as to avoid violations of first-order stochastic dominance, and yet accounts for some of the challenges presented above, especially common ratio effects. The formal proximity with EU has been clarified through axiomatizations of RDU, which shares the EU axioms *but with a weakened form of independence* (see Quiggin 1982; Quiggin and Wakker 1994). Its explanatory power has been offered *p*-&-*P*sychological interpretations in terms of optimism and pessimism, which admit normative interpretations unlike with PT (see Diecidue and Wakker 2001). However, if RDU makes a step toward psychologists' probabilistic sensitivity, it still remains silent regarding the specific perspective taken by the latter, especially Kahneman and Tversky, on consequences, i.e., the

[17] 'RDU' is now preferred to 'RDEU' roughly because, as Wakker puts it in a text about these two expressions, 'expected' does not add much information and "[i]t is very very important that terms be short and tractable" (see: http://people.few.eur.nl/wakker/miscella/rduquiggin.txt ; last consulted 03/05/2016). I thank him for having pointed this text to me.

[18] It can be illustrated briefly. The main specificity of Quiggin's decision weights shows up for lotteries that have at least three non-null consequences, and so it is not appropriate to compare it directly with Kahneman and Tversky's (1979) original approach (see Wakker 2010, pp.274-5 on this point). The distinction between both approaches is that Quiggin's weighting function is not applied to consequences' probabilities, but to ranks, i.e., to cumulative probabilities. Consequences need to be already *ranked*, i.e., the subscripts of the xs are meaningful, e.g., $x_1 \geq ... \geq x_n$. Decision weights for a given consequence are derived from the entire probability distribution of consequences. More precisely, they are derived from the *ranks* of consequences x_i, which are the probability of receiving a given consequence *or* anything better, i.e., the *cumulative probability* $p_i + ... + p_n$, not just p_i. Take the lottery (borrowed from Wakker 2010, pp.158-9) where you have $\frac{1}{6}$ chance of winning €80, $\frac{1}{2}$ chance of winning €30, $\frac{1}{3}$ chance of winning €20, i.e., $X = (x_1, p_1; x_2, p_2; x_3, p_3) = (€80, \frac{1}{6}; €30, \frac{1}{2}€20, \frac{1}{3})$. There are four ranks: 0, $\frac{1}{6}$, $\frac{2}{3}$ and 1, corresponding respectively to the cumulative probabilities of more than €80 (0), €30 $(0 + \frac{1}{6})$, €20 $(0 + \frac{1}{6} + \frac{1}{2})$ and €0 $(0 + \frac{1}{6} + \frac{1}{2} + \frac{1}{3})$. Probabilities of individual consequences can be expressed in terms of *differences between ranks*: the probability of 'winning €80' is $\frac{1}{6}$ which is equal to the rank of winning more than €30 $(\frac{1}{6})$ minus the rank of winning more than €80 (0); by the same reasoning, the probability of 'winning €30' is $\frac{1}{2}$ (i.e., $\frac{2}{3} - \frac{1}{6}$) and the probability of 'winning €20' is $\frac{1}{3}$ (i.e., $1 - \frac{2}{3}$). Finally, the decision weight w_i of a consequence x_i is the difference between ranks transformed by a weighting function $\pi(.)$: $U(X) = u(x_1)w_1 + u(x_2)w_2 + u(x_3)w_3 = u(€80) \times [\pi(\frac{1}{6}) - \pi(0)] + u(€30) \times [\pi(\frac{2}{3}) - \pi(\frac{1}{6})] + u(€20) \times [\pi(1) - \pi(\frac{2}{3})]$.

central distinction between gains and losses. In short, RDU cannot straightforwardly account for sign effects. The critical role played by RDU in the post-1979 developments around the making of behavioral economics is the adoption of this approach to decision weights by Tversky and Kahneman (1992) themselves in CPT. Hence CPT avoids violations of first-order stochastic dominance, and extends Quiggin's approach to the asymmetry between gains and losses, notably through two weighting functions, one for positive consequences and one for negative consequences (see Tversky and Kahneman 1992, fig.3). CPT therefore makes new empirical predictions (see Fennema and Wakker 1997; or Wakker 2010 appendix 9.8 for comparisons) and has also been axiomatized (see Wakker and Tversky 1993; Chateauneuf and Wakker 1999). This opened the door to some possible reconciliations between behavioral and standard economics, at least for decision under risk. Indeed, since the publication of CPT, there have been much discussions over the appropriate functional form of the weighting function, the implications of its parameter values, and other quantitative issues into which this dissertation will not delve (see Fehr-Duda and Epper 2012).

One such quantitative issue is however worth briefly discussing, if only because, as we shall see, it indirectly underlies some normative contentions between standard and behavioral economics. Recall that, informally and atheoretically, there are three basic risk attitudes that economists try to empirically measure: *risk neutrality* if the decision maker is indifferent between a lottery and the certainty of its expected value, e.g., '50% chance to get €6000 and 50% CHANCE to get €0' ∼ 'The CERTAINTY of getting €3000'; *risk seekingness* if he prefers the lottery to the certainty of the expected value; and *risk aversion* if he prefers the certainty of the lottery's expected value to the lottery. Economists' consequentialist sensitivity manifests itself in their uses of EU for measuring risk preferences by deriving indexes of risk aversion which only take into account the decision makers' attitudes toward consequences, not toward their probabilities. Indeed, these measures are derived from the theoretical convention that the whole of risk attitudes are identified with measures of the curvature of the subutility function (risk neutrality corresponds to $u(.)$ being linear, risk seekingness to $u(.)$ being convex and risk aversion to $u(.)$ being concave). Under either PT or CPT, the use of a weighting function implies (also under RDU) that risk attitudes are not anymore taken to be *entirely* determined by the curvature of the subutility function, but *also* relates to the curvature of the weighting function – and to the

curvature of loss aversion (see the preceding chapter, and below). Thus, Wakker (2010) insists at length that all the sophisticated measures of the curvature of utility developed in standard economics should not anymore be called measures of risk aversion in non-EU settings, they should just be called what they are, i.e., measures of utility curvature.[19]

The issues underlying the value function of PT and CPT touch mainly the economists' consequentialist sensitivity on the way of dealing with asymmetries between gains and losses, i.e., with sign effects. The value function handles these asymmetries through 'reference-dependence' as explained in the previous chapter, i.e., consequences are taken as gains or losses with respect to a reference point: it is $u(x)$ or $u(-x)$). EU, on the other hand, handles them through 'asset integration', by counting as a consequence *not* the consequence of a given choice *per se* but its consequence on the decision maker's final wealth through its initial wealth w: it is $u(w+x)$ or $u(w-x)$). The most salient contention between these two approaches arose with respect to a quantitative challenge, the so-called 'Rabin paradox'. Rabin (2000) considers the following preference:

A: nothing
B: 50% CHANCE of winning €110 and 50% CHANCE of losing €100

Through some algebraic manipulations, he shows that absurd quantitative measures of risk aversion are implied by EU with asset integration, regardless of both initial wealth levels and functional forms except for concavity. In particular, the previous preference for nothing implies the following one:

$\mathbf{A}'$: nothing
B': 50% CHANCE of losing €1000 and 50% CHANCE of winning any sum of money

In short, plausible risk aversion over win/lose small stakes imply implausible risk aversion over large ones (e.g., replace 'any sum of money' by $1.000.000). This comes from the whole of risk attitudes being derived from the marginal utility of money in EU. Rabin's solution is

[19] For more on the quantitative empirical issues underlying CPT *versus* EU, see Harrison and Rutström (2008) and Holt and Laury (2014). The theoretical background needed to understand these issues is pedagocially displayed by Mas-Collel et al. (1995, chap.6), Varian (2005, chap.12), Baccelli (2016) and especially Wakker (2010, chap.3); see Gollier (2001) and Meyer (2014) for more advanced discussions.

to drop asset integration and to use loss aversion, which implies reference-dependence, which is tantamount to using PT's value function. Regarding the purpose of this chapter, there are at least three points of contentions between standard and behavioral economists over the use of that function.[20]

Firstly, as for probability weighting, loss aversion is not necessarily considered as irrational in so far as it can be thought of as decision weights whereby the decision maker consciously put twice more weights on losses than on gains (Wakker 2010, p.239). This underlies the most frequent criticism of prospect theory: there is only sporadic considerations but no systematic theory of how the reference point changes, which have to be guessed in a pragmatic way by the decision theorist (see, e.g., Wilkinson and Klaes 2012, p.190; Wakker, 2010, p.241-2 and sect.8.7; Stommel 2013). Thus, there is a lack of systematicity not only about the fixation of the reference point, but also, by implication, on the underlying (ir)rationality of the behavior under analysis. In other word, the point made in the previous chapter about the situation-dependency of rationality judgments made from PT's value function under certainty carries over under uncertainty.

Secondly, there is a third way of resolving the asset integration *versus* asset isolation conflicts, by dropping neither of them but making some value judgments on the content of decision makers' preferences. This is the underlying normative contentions that was mentioned above as being indirectly connected with the quantitative challenges posed by behavioral economics. This is for instance Wakker's position (2010, chap.8), who argues that rationality requires risk neutrality (or very weak risk aversion) over small consequences. In other words, Wakker argues against the translation of consumer sovereignty in decision under uncertainty, i.e., that the content of preferences (for or against risk) is amenable to rational criticism.

Finally, the real difference between standard analysis and prospect theoretic analysis is not the notion of reference-point *per se*, but the fact that it can be variable in the latter and is necessarily fixed in the former. Indeed, EU's asset integration and prospect theory's reference-

[20] For more on the Rabin paradox, see Rabin and Thaler (2001) and Barberis, Huang and Thaler (2006) from behavioral economics' perspective and the references discussed by Wakker (2010, pp.244-5) from standard economics' perspective. There has always been a tradition in standard economics (esp. in experimental economics) to model *EU over income* rather than over final wealth. Many response to the Rabin paradox used EU over income, but without remarking that this is tantamount to reference-dependence (see Harrison, Lau and Rutström's 2007 appendix; and Wakker 2010, p.245). Regarding the positive/normative issue, because Rabin's solution is through loss aversion which entails reference-dependence, the only remaining disagreement between users of EU over income and prospect theory seems to be about the (ir)rationality of decision makers.

dependence are in agreement when the latter's reference point is a fixed level of initial wealth, so that there is "a unique relation between the [consequences] and the final wealth" (Wakker 2010, p.238); this unique relation is simply initial wealth plus consequences equals final wealth. It is when, for a given decision maker and a given set of choices, this unique relation is broken *during the analysis* by a change in reference point that both approaches disagree. In such cases, the traditional behavioral economics position of keeping the standard approach as the normative benchmark and using psychologists' theories as positive models surfaces rather clearly: "traditional EU is, in my opinion, the hallmark of rationality, any deviation from final wealth due to reference dependence is utterly irrational" (Wakker 2010, p.245). In short, "[e]xpected utility is the right way to make decisions" (Thaler 2015, p.30).[21]

Over time

We now turn to time preferences, by way of contrast with what has been discussed for risk preferences. Here, the main inspiration from *P*sychology comes from psychiatrist George Ainslie (1975; 1992) instead of Kahneman and Tversky, whose theory of hyperbolic discounting played a role somewhat analogous to PT in the making of behavioral economics. In 1975, Ainslie published a famous paper putting together a large amount of empirical observations and theoretical considerations from various disciplines on human and animal individual behaviors over time. There and in subsequent publications, he has argued for the high regularity of two phenomena: "impulsiveness", responsible for dynamic inconsistencies and "impulse control", i.e., various ways by which humans and animals try to avoid dynamic inconsistencies. It can be argued that one parallel to the consequentialist *versus* probabilistic sensitivities that divided

[21] It should be noted that there have been a few attempts at endogneizing changes of reference point, notably by Botond Köszegi and Rabin (2006, 2007, 2009) in several settings (see also Schmidt 2003; Schmidt, Starmer and Sugden 2008). Köszegi and Rabin formally characterize reference point as "probabilistic beliefs about the relevant consumption outcome held between the time she first focused on the decision determining the outcome and shortly before consumption occurs" (2006, p.1141). Their determination of reference-point is (1) counterfactual (see e.g., 2006, fn3 p.1138), (2) temporal, and (3) rational. The last feature is surprising regarding what has just been discussed. They assume rational beliefs (i.e., rational expectation) for simplicity and often stress (esp. in the 2006 and 2007 conclusions) the inability of their models to account for some irrational behaviors demonstrated in behavioral economics (though see 2009, sect. III, where some of these irrational behaviors are seen as rational through the lenses of their model). Their main formal innovation in decision making under risk is their uses of tools from game theory as in models of time preferences (see below), i.e., selves playing a game, the equilibrium of which determine the utility, preference and behavior of the decision maker (see Köszegi 2010). Thus, there is a (not so) implicit interaction with the dimension of time, in line with the developments on the primacy of risk preferences discussed in the previous subsection.

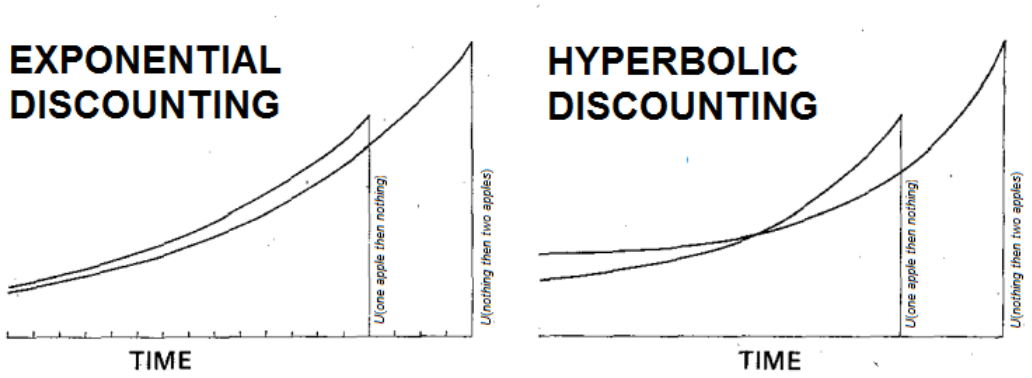

Figure 2.2: Exponential versus hyperbolic discounting and the apple example

standard economists and psychologists in the dimension of uncertainty is, in the dimension of time, the distinction between impatience and impulsivity.

We can illustrate that through Thaler's (1981) apple example. The exponentially discounted utility of 'one apple early and then nothing' should always be superior *or* inferior to the the exponentially discounted utility of 'nothing and then two apples later'. There cannot be preference reversals and this absence of dynamic inconsistencies is due to the rate of discounting being a positive constant which is usually interpreted as *impatience*. By contrast, the rate of discounting is not constant in an hyperbolic discount function. The closer the time of evaluation is to the time when the consequence happens, the greater the rate of discounting which Ainslie interprets as impatience *and impulsiveness*. Hence, the hyperbolically discounted utility of 'nothing and then two apples' can be superior to the one of 'one apple early and then nothing', but the closer the time of evaluation is to the time of the early apple, the faster its utility increases compared to the increase of the utility of the two later apples. Figure 2.2 provides a graphical comparison between exponential and hyperbolic discounting to capture the contrast just made (through a modification on Ainslie 1975, fig.1).

Under exponential discounting, a choice for 'one apple early and then nothing' should be represented by its utility being always superior to the utility of 'nothing and then two apples later'. This contradicts the outcome of Thaler's example. In line with it, however, under hyperbolic discounting the utility of 'nothing and then two apples later' is represented as being superior

to the utility of 'one apple early and then nothing' at early evaluation times before smoothly reversing as evaluation times get closer to the time of getting the early apple. Roughly, according to Ainslie, impulsivity is responsible for the crossing of the curves, and the combination of impatience, impulsivity and impulse control is responsible for their smoothness.

The most famous introduction of Ainslie's work in economics was made by behavioral economist David Laibson (1994; 1997) in his models of quasi-hyperbolic discounting. The latter can be seen as introducing the consequences of hyperbolic discounting within the formalism of exponential discounting. The quasi-hyperbolic discount function in Laibson's models is a step function. It discounts the utility of a plan as its exponentially discounted utility with an additional discounting parameter β that is (i) necessarily equal to one at the time of evaluation and (ii) can then take any value between zero and one to multiply all exponentially discounted subutilities. Formally, this is represented by $U^t(X) = u(x_t) + \beta \sum_{i=1}^{\infty} \delta^i u(x_{t+i})$, i.e., the utility of a plan is the undiscounted subutility of the consequence that happens at the time of evaluation (as in standard exponential models) added to the sum of the exponentially discounted subutilities each multiplied by the parameter β. It is easy to see that if $\beta = 1$ even after t, quasi-hyperbolic discounting reduces to exponential discounting and thus implies dynamic consistency. It is also easy to see that if $\beta = 0$ after t, the decision maker does not care at all about LATER and THE FUTURE in the sense that all consequences after the date of evaluation contribute nothing to the utility of the plan under evaluation, which is thus just the subutility of the INSTANTANEOUS consequence (i.e., $U^t(X) = u(x_t)$). Thus, in-between, i.e., for $0 < \beta < 1$, the "perturbation to the "standard" [time] preferences" (Laibson 1994, p.14) triggered by β are so that the closer β gets to 0, the greater the evaluation of the plan is biased by the PRESENT consequence (with respect to an evaluation under exponential discounting). Hence the two alternative labels for quasi-hyperbolic preferences in the literature, 'present-biased preferences' or more neutrally 'β-δ preferences'.[22]

The quasi-hyperbolic functional form is not directly Laibson's contribution. It was originally developed by Edmund Phelps and Robert Pollak (1968) to model intergenerational altruism in macroeconomics. In their model, the PRESENT generation is denoted by t, and is evaluating a

[22] Some precedents to Laibson's introduction of Ainslie's work in behavioral economics should be acknowledged, especially Thaler (1981) from an empirical perspective and, from a theoretical perspective, Thaler and Shefrin (1981) (discussed in the next chapter) and Loewenstein and Prelec (1992).

plan where the consequences are its OWN consumption (x_t), along with the consumptions of all FUTURE generations (x_{t+i}). The latter's utilities are discounted exponentially (through δ^i) and all multiplied by the constant β which measures "the degree to which the present generation values other peoples' consumption relative to their own", with $\beta = 1$ representing "perfect altruism" and $0 < \beta < 1$ "imperfect altruism" (p.186). This functional form, used in a dynamic game theoretical setting, enables Phelps and Pollak to characterize equilibria corresponding to more or less optimal saving rates depending on the degree of altruism. Laibson's main interpretational move is to turn Phelps and Pollak's generations into "temporal selves" (Laibson 1994, p.13), i.e., into a so-called *multiple selves model*. Thus, intertemporal decision problems are modeled by indexing by t the PRESENT self who is currently evaluating a plan ($U^t(X)$) with respect to his own undiscounted utility ($u(x_t)$) and the exponentially discounted utilities of the FUTURE selves ($\sum_{i=1}^{\infty} \delta^i u(x_{t+i})$), anticipating that they will be biased towards THEIR PRESENT when they will re-evaluate the same plan (β). Regarding the theme of this chapter, notice how the formal similarities in two of the three dimensions allowed Laibson to use a functional form originally designed for the social preferences of generations to model the time preferences of a single individual.[23]

At this individual level, quasi-hyperbolic discounting can, like hyperbolic discounting, imply the preference reversals that characterize dynamic inconsistency. But unlike with hyperbolic discounting, the reversal is not smooth. Before the time of evaluation coincides with the occurrence of consequences, quasi-hyperbolic discounting is like exponential discounting but 'restrained' by $\beta < 1$. When the time of evaluation coincides with the occurrence of consequences, then $\beta = 1$, which suddenly makes their utility higher, especially compared to the utility of the consequences occurring later. Figure 2.3 provides a graphical representation of the explanation just made (again with the apple example and through a modification on Ainslie 2012, fig.3).

If time is discrete, then kinks due to β suddenly becoming equal to 1 should be interpreted as the beginning of the evaluation period where consequences (either one or two apples) are occurring.

[23] Elster (1979, p.71ff) should be acknowledged as a precedent for the intuition of adapting Phelps and Pollak's model at the individual level.

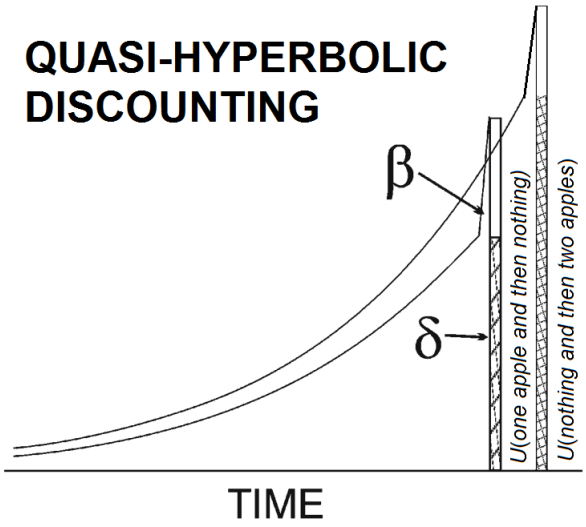

Figure 2.3: Quasi-hyperbolic discounting and the apple example

By contrast with the discussion of risk preferences above, there are some contentions between behavioral economics and *P*sychology, i.e., hyperbolic discounting *versus* quasi-hyperbolic discounting. That is, the issue of interdisciplinarity between economics and *P*sychology is more internal to behavioral economics than it was for risk preferences. If quasi-hyperbolic discounting has been widely used in behavioral economics, it is mainly for the good compromise it offers between the analytical tractability of its exponential discounting structure and the *p*sychological realism of its implied preference reversals (the most important qualitative property of hyperbolic discounting from a revealed preference perspective). Reviewing the mathematical structure of much of the discount functions in economics and *P*sychology, psychologist John Doyle (2013, p.125) remarks that despite the quasi-hyperbolic discontinuity being motivated primarily by analytical matters, it is "a model devised by economists for economists [which] has the capacity to capture a psychological distinction that the other models cannot", namely the "*qualitative* difference between now and any time to come, which we all appreciate intuitively, and which language mirrors in verb tenses". Doyle further argues that this qualitative difference is all the more important to be captured because it also occurs in other situations, so that quasi-hyperbolic models are useful as partly able to capture the facts that "[r]ewards that are *certain*,

mine and *now*", i.e., consequences in the Rastier's identity zones of the three dimensions, "are all over-valued" (ibid). This remark is obviously well in line with the perspective of this chapter. Nevertheless, this sympathetic comment from a psychologist should be contrasted with the unsympathetic ones recently addressed by Ainslie (2012) to quasi-hyperbolic discounting. Among the three main criticisms from Ainslie (2012) to be discussed, the first two are worth mentioning because the underlying issues will resurface later; only the last one is central for the purpose of this section.

Firstly, Ainslie claims that all the sophisticated techniques of self-control that he has observed and theorized through the years cannot be captured in a model where the whole of the preference reversal comes from the anticipation of a 'last-minute' sudden "β spike" of utility for a given consequence (the expression is from Ainslie 2012). In other words, Ainslie argues from *P*sychology that decision makers exercising self-control anticipate a much more complex process. This claim is central for the current debates between *standard* and behavioral economics around quasi-hyperbolic discounting, which are discussed in the next chapter (see the discussion of the work of Gul and Pesendorfer in §3.2.2).

Secondly, Ainslie argues that because the shape of β-δ curves in the long-run are similar to the shape of exponential curves, they have just as much trouble as the latter notoriously has in explaining the great diversity in empirical measures of discount rates. Recall that exponential discounting implies that in each period the utility of the next period's consequences are discounted by the same factor, namely $\frac{1}{1+\rho}$, where ρ is a constant discount rate. Thus, the utility of consequences between two periods in the future are discounted at the same rate than between two periods in the near present (or in the very distant future for that matter). Though the underlying methodological issues are numerous, studies that try to elicit explicit or implicit discount rates report a very wide variety of values, from negative rates to infinitely positive ones, even (and this is the most crucial finding for the present argument) within a same study for a same individual (see Frederick, Loewenstein and O'Donoghue 2002, table 1; Wilkinson and Klaes 2012, chaps.7-8). The general pattern seems to be that discount rates decline over time, though there are many exceptions. What Ainslie argues is that β-δ curves reproduce this qualitative pattern but not quantitatively as well as hyperbolic discount curves, and that there is no way to naturally explain the exceptions as hyperbolic curves does. By contrast with preference

reversals violating dynamic consistency, this is a quantitative empirical issue akin to the ones on the measurement of risk attitudes mentioned above.

Thirdly, Ainslie's own analysis of hyperbolic discounting makes use of dynamic game theory in an informal manner to cast most of his theoretical explanations (see esp. 1992). Another difference (besides impatience *versus* impulsivity) that can characterize contrasting sensitivities between economists and psychologists on time preferences is the respectively bad and good epistemic values attached to the multiple equilibria resulting from both hyperbolic and quasi-hyperbolic models. As Ainslie puts it, "[i]nformal discussion has suggested that the obstacle [in the adoption of such models by economists] is economists' determination to predict unique decisions, in principle at least, from a given set of prior motives", though it may be that "a technically chaotic mechanism [...] is the way of the world" (2012, p.21 and p.29). In this respect, it is very illustrative to note the following evolution in Laibson's work. At the beginning of his dissertation, Laibson states that "[t]he primary goal of this thesis is to formalize and extend Ainslie's psychological analysis, by explicitly modeling individual decision-making as an intra-personal game" (1994, p.10). His interpretation of the multiple equilibria result is that this is not necessarily a bad thing because it could be interpreted that "the model generates heterogeneous behavior without making recourse to heterogeneous preferences" (1994, p.40). Furthermore, with the same preferences, two decision makers may end up in equilibria that are Pareto inferior or superior to one another, which Laibson interprets as decision makers with same preferences but self-acknowledged "bad habits" ending in Pareto-inferior outcomes or "self-control" ending in Pareto-superior outcomes (ibid). This reasoning is indeed close to Ainslie's work. However, not only does this interpretation disappears in the 1997 paper, but he has concentrated some efforts from the 2000s onwards with Christopher Harris to find a way of getting a unique equilibrium with β-δ preferences (see Harris and Laibson 2013 for the first published version).

A last comment on the theoretical consequence of time preferences seems in order regarding the perspective of this chapter: the asymmetry between gains and losses at play in sign effect has attracted neither the theoretical attention nor the controversies they have in the risk preference literature. Yet, an early introduction of hyperbolic (*not* quasi-hyperbolic) discounting in economics was made by Loewenstein and Prelec, in several joint and individual papers, with specific theoretical attention to sign effects. Their formal contribution (see esp. Loewenstein

and Prelec 1992) consists in a generalized hyperbola that nests exponential, quasi-hyperbolic and hyperbolic discounting function as special cases. Loewenstein and Prelec's way of introducing hyperbolic discounting in economics is explicitly inspired from Kahneman and Tversky's prospect theory, especially by making the notion of temporal reference-dependence central in their reflections (which implied working with different objects of choice, "*temporal prospects*" as adjustment to consumption rather than consumption *per se*). This line of formal unification has however not been pursued in behavioral economics, which could have taken the three dimensions altogether as they remark at some point that "interpersonal comparisons of material outcomes reflect a concern both for absolute and relative differences between one's own and other persons' payoffs" (Prelec and Loewenstein 1991, p.784). Though this very brief comparison to social preferences is the only point at which they go into the three dimensions, it is worth keeping in mind that the notion of reference-dependence has that kind of unifying potential in behavioral economics.

What about the positive/normative issue with respect to time preferences? In this dimension, the contentions are rather different than the ones we have seen for risk preferences. There might have been some ambiguities between the positive/normative distinction at the individual level and at the level of subareas in economics (i.e., positive versus normative economics, see the previous chapter). Samuelson's (1937) original comments at the individual level are at best agnostic: he stresses the arbitrariness of the functional form throughout and only highlights the rationality underlying dynamic consistency for the theoretical case of perfect markets with interest rate. At the level of subareas in economics, he is more explicit on exponential discounting having no place in normative economics: "any connection between utility as discussed here and any welfare concept is disavowed [...] [and] any influence upon ethical judgments of policy is one which deserves the impatience of modern economists" (p.161). Some have interpreted the first part of this quote as belonging to the individual level of the positive/normative issue (e.g., Frederick et al. 2002, p.355). From what has been argued in the previous chapter, this would not be a fallacy in itself (i.e., there are several methodological arguments to connect both levels). However, it goes against the general tendency shared by both standard and behavioral economists to derive value judgment of rationality from exponential discounting because it prevents intertemporal preference reversals (see, e.g., Gollier 2001; Loewenstein et al. 2015,

pp.61-65, see next chapter §3.3.2). And symmetrically, value judgments of irrationality are derived from quasi-hyperbolic discounting because it implies such reversals. On this latter point, there are some subtleties that hinge on the use of quasi-hyperbolic discounting to model "sophisticated" *versus* "naive" decision makers. Recall that in Laibson's models β captures the evaluation of a plan by a self at time t who *anticipates* all the next $t+1$ selves to be biased towards THEIR PRESENT if re-evaluating the same plan. If these anticipations are assumed to be correct, then the multiple selves model is taken to be about a sophisticated decision maker; if incorrect, it is about a naive one. The normative issues underlying time preferences are just stated here, but will be discussed in more depth in the next subsection and in the next chapter.[24]

Regarding other people

Turning to social preferences, there are four main theoretical papers seeking a unifying explanation for the empirical regularities observed in games such as the dictator game (Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Andreoni and Miller 2002; Charness and Rabin 2002; the latter can be seen as an extension of Rabin 1993, which has the same motivation as the others). There are two main differences with the contributions on risk and time preferences. On the one hand, there is not one but *several* alternatives (at least five) to the standard model (which, moreover, has always had controversial interpretations issues of self-interest and altruism). On the other hand, though some inspirations are drawn here and there from social psychology, there is no a substantial influence from one psychologist (or even from a well-identified set of psychologists for that matter). For these and other reasons the presentation of alternative*s* to self-centeredness here will not be illustrated with notations pertaining to their uses of formal language (as was done with risk and time preferences).[25]

What is worth discussing informally is how they all introduce non-self-centered motives in utility functions, which can be clearly interpreted as specific cases of Sen's sympathy, i.e.,

[24] Much of Ainslie's work focus on self-control and the normative medical implications of pathological self-control problems; the complex implications for standard and behavioral economics of the normal/pathological categories will not be discussed in this dissertation (see Vallois 2011; 2012a; b; 2014).

[25] The other reasons are because the theoretical explanations discussed here are not entirely focused on the dictator game and not at all concerned with the types of regularities depicted in the first subsection (e.g., on 'social distance'). Furthermore, unlike the respective alternatives to expected utility and exponentially discounted utility models, there is not a unique theoretical alternative here, and illustrating the differences among the formal structure of models of social preferences would take too much space (and this has already been done, see esp. Camerer 2003, chap.3; Sobel 2005, sect.3; Fehr and Schmidt 2006, sect.3; Wilkinson and Klaes 2012, sects.10.5-6).

different ways by which the decision maker's utility is dependent on "the state of others" (2002, chap.1, p.35). For a given allocation, such dependence can be (1) on whether THE OTHER gets more or less than the decision maker (Fehr and Schmidt 1999), (2) on whether THE OTHERS get on average more or less than the decision maker (Bolton and Ockenfels 2000), (3) on THE OTHER(S) who get(s) the least, or on (4) how much WE get in total (both Andreoni and Miller 2002 and Charness and Rabin 2002 propose characterizations for the two latter cases). Notice these correspond to positions in ethics, moral, or political philosophy: (1) and (2) are two species of egalitarianism and (3)/(4) represent the Rawlsian/utilitarianist contrast. Though not well illustrated by the simple choice over an allocation, the dependence of the decision maker's utility can also be on (5) THE OTHER's (a) own motives and (b) beliefs about the decision maker's motives (Rabin 1993). Notice that (5) does not match a position in ethics, moral, or political philosophy; it corresponds to the realm of so-called 'psychological game theory' (see Dufwenberg 2008). In terms of Sen's conceptual apparatus, (1)-(5) violate the requirement of self-welfare goal (i.e., that own welfare maximization is the unique goal) because they all imply that the decision maker is not *only* maximizing his own welfare, though they all imply that he is nevertheless *also* maximizing his own welfare. Wilkinson and Klaes (2012, p.422) make a brief remark regarding Sen's reflections and the social preferences literature around behavioral economics: the possibility in Sen's conceptual apparatus of violating self-welfare goal by the disappearance of own welfare maximization as a goal (even if compensated by the addition of others' welfare maximization as another goal) "is alien to economics, since it cannot be incorporated into any utility maximization model, whether standard or behavioral". Furthermore, by contrast with the main interpretations of violations of the independence axiom and dynamic consistency by behavioral economists (cf. above), (1)-(5) are always accompanied with statements about the rationality of the underlying behaviors. This *value* judgment comes from the theoretical *conventions* of using formal language to construct (just like in standard economics) well-behaved utility functions from the maximization of which game-theoretical equilibria can be derived to explain and predict *factual* behaviors in experimental games (see esp. Andreoni and Miller 2002; and Fehr and Schmidt 2006, sect.2.2).

All (1)-(5) insist that one of the contribution of their formal model is the possibility of capturing the value judgment of the decision makers, "the concerns for fairness and equity *of*

the economic actors being studied" without "incorporat[ing] economists' judgments of fairness and equity" as usually done in normative economics (Rabin 1993, p.1282, my emphasis). As Wilkinson and Klaes (2012, p.396) put it, they take "the view that 'fairness is in the eye of the beholder"'. In short, the use of formal language by the economist is supposed to capture how decision makers would justify their behaviors in ordinary language, though without talking to them. Notice that 'fairness is in the eye of the beholder' could be interpreted as guiding actions independently of the decision maker's welfare, i.e., as cases of Sen's commitment – "breaking the tight link between individual welfare (with or without sympathy) and the choice of action" (2002, ibid). But this is not plausible here because (1)-(5) are all meant to capture "how people's attitudes toward fairness influence their behavior and well-being" (Rabin 1993, ibid), through the uses of formal language in the maximization of utility and the derivation of game-theoretic equilibria. But the empirical regularities that are the main motivation for (1)-(5) are confined to lab experiments with designs as 'neutral' as possible compared with the situational factors related to social distance discussed in the first subsection. Hence, it can be argued that the impossibility for (1)-(5) to capture violations of self-welfare goal or commitment (à la Sen) does not constitute a criticism of their lack of explanatory power or of their *p*sychological realism *per se*. Such criticism could however be made, in line with Sen's original criticisms, to explanations of real-world phenomena extrapolated qualitatively from (1)-(5), which are made here and there in the papers referenced above (see also Meier 2007). We shall not conduct such a critical exercise here, but notice that because (1)-(5) capture various degrees of Sen's sympathy, it is at some distance from the decision maker as "close to being a social moron" (Sen 1977b, p.336) that was the original target of Sen. Thus the critical exercise may not be as radical as one might think; it would mainly consist in discriminating sympathy from commitment in the field, thus showing what remains the be incorporated in models of social preferences (or, depending on one's point of view, what remains beyond their scope).

Although (1)-(5) are primarily motivated by experimental results in 'neutral' situation, it is often remarked that including "specific economic variables like rules of the game, as well as social variables like the level of anonymity, the sex of one's opponent, or the framing of the decision" (Andreoni and Miller 2002, p.738) is in the agenda for future theoretical developments. It is also not rare to read that the notion of identity (i.e, the decision makers' personal, social

and/or political identities) should be the central notion through which to achieve such theoretical unification (e.g., Camerer 2003, p.476; see also Fehr and Shmidt 2006). Arguably, this is a project that has been conducted in the influential work of Roland Bénabou and Jean Tirole (2011) which extends earlier reflections of these authors on non-monetary incentives and self-control to social preferences as understood in (1)-(5). For the purpose of this subsection, three broad aspects of Bénabou and Tirole's work are worth commenting. Firstly, by contrast with (1)-(5), Bénabou and Tirole do not emphasize that the behaviors they are modeling are necessarily rational (though they do not pronounce *explicit* value judgments of irrationality, some are obviously latent in their work). Secondly, as in (1)-(5), references from social *P*sychology are only briefly mentioned and there is no focal psychologist (in the sense in which Kahneman and Tversky and Ainslie are respectively focal for risk and time preferences). But unlike in (1)-(5) Bénabou and Tirole (2011) mention much more of such references and seek to show how various phenomena observed by social psychologists can be captured in the formal language of their model, which at the same time highlights their economic implications. Finally, the third aspect is related to the central role of the notion of identity in these interpretations and economic implications. At one point, they argue that Sen's commitment, especially when explained by decision makers' identities, "can be formalized, in a way consistent with consequentialist rationality" through their model (Bénabou and Tirole 2011, fn32 pp.828-9). Hence the only way Sen's commitment can be integrated into the social preferences literature around behavioral economics is if it is devoided from one of its initial ambition, namely to capture some non-consequentialist form of rationality relevant for the behavioral foundations of economics.

Summing up, this subsection tried to characterize the contrasts among the three dimensions of the issue of interdisciplinarity as well as the positive/normative issue. Roughly, for the issue of interdisciplinarity, the main contrast is that while there are well-identified work in *P*sychology influencing behavioral economics' models in the dimension of uncertainty and time, this is not the case in the dimension of other people. For the positive/normative issue, while the standard model *versus* behavioral models opposition tends to map rather clearly the value judgments of rationality *versus* irrationality derived from them in the dimension of time, and with more qualifications in the dimension of uncertainty, there are no such clear relations in the dimension of other people.

From a constructive perspective intended to motivate further empirical and theoretical contributions, the next subsection proposes a critical assessment of the standard and behavioral contributions discussed so far in the three dimensions.

2.1.4 The experimental translation of rationality, narrativity and identity

In the previous chapter, I tried to make the case that arguments against the methodological convention of 'given preferences' in discussions of economic behavior under certainty "inevitably drift toward one of the three dimensions of economic behaviors and take this dimension to be a *primary* determinant of individual preferences" (in §1.3). In this sense of 'primacy', it was then argued that there is a 'primacy of risk, time and social preferences' illustrated by the role of reflective reasoning, of hedonic experiences and of ethical reasoning in the works of Tversky, Kahneman and Sen, respectively. The goal of this subsection is to provide a critical assessment of the contributions discussed in the three previous subsections from an historical, methodological and philosophical perspective on these three primacies. Following Mongin (2014) we argue for a revival of what he calls "*the experimental translation of rationality*", a tradition that stems from the work of Maurice Allais and to which reflective reasoning in the work of Tversky belongs. The work of Robert Strotz, who played a somewhat analogous role on the behavioral economics of time preferences to the one Allais played for risk preferences, is discussed to flesh out the normative strength of dynamic consistently and show that it is amenable to the same criticisms addressed to Kahneman's primacy of time preference regarding the role of *narrativity* in individual behavior. Finally, the role of *identity* in the formation of commitment within Sen's rationality as reasoned scrutiny is briefly explained.

The common ratio effect discussed in this section is the generalization of a decision problem originally devised by Allais and other French economists (see Pradier and Jallais 2005), along with other decision problems that are today called the 'Allais paradoxes'. The initial intention of Allais and his collaborators was to show that EU and its independence axiom were problematic on *both* positive *and* normative ground. As emphasized by Mongin (2014; see also Guala 2000 and Heukelom 2014), Kahneman and Tversky (1979) were explicitly inspired by the Allais paradoxes in their positive criticism of EU. But like the behavioral economists that followed them, they did not side with Allais on the normative dimension. Hence, it is worth refreshing memories

regarding how these normative issues have been raised, and to propose some developments to revive them.[26]

Allais (1953, p.518) contends that it makes no sense to define rationality by "direct" reference to the EU's additive formulation or by "indirect" reference to the axiom of preferences implying it. In his positive and normative uses of counter-examples (i.e., the Allais paradoxes), Allais (1953; 1979a;b) defends an alternative conception of rationality, that he labels "experimental" (1953, p.521, my translation) or "pragmatic" (1979b, p.467): "it is possible to define rationality [...] by observing the acts of people *of whom there exist independent grounds, i.e. without recourse to any consideration of random choices, for believing that they behave rationally*" (ibid). More precisely, Allais' conception of rationality concerns the acts of those people who are acquainted with uses of formal language, i.e., that are "highly conversant with the theory of probabilities, having generally extensive mathematical knowledge" (ibid, p.468).

Who are these people? They are the decision makers Allais recruits to answer his choice problems. Besides high-ranking officials from various governmental institutions (see Pradier and Jallais 2005), this notably includes decision theorists. Even the decision theorists defending EU are decision makers in Allais' choice problems. This is well known through the story of Leonard Savage, a prominent contributor and defender of EU who violated his own independence axiom. Savage (1972 [1954], p.103) famously argued that *after careful reflections*, his choices were irrational so he would change them to obey his axiom if given the opportunity, which made the models within EU normative on his account. But the choice problems constructed by Allais were also intended as normative criticism of EU. A key feature of Allais' normative argument is that EU does not properly account for the specificity of Certainty in decision making under Risk, especially regarding the *values* of prudence, security, and the avoidance of ruin, that rationalize the *factual* violations of EU's *conventions* by the rational men he selected (see Allais 1953; 1979b; Guala 2000; Pradier and Jallais 2005; Heukelom 2014; and esp. Mongin

[26]It is worth contrasting Kahneman and Tversky's with Quiggin's positions on this issue. For the former, patterns of preferences underlying common ratio effects among others are "*anomalies* implied by PT", and "[t]hese departures from expected utility must lead to normatively unacceptable consequences [...] normally corrected by the decision maker when he realizes that his preferences are inconsistent" (Kahneman and Tversky 1979, p.277, my emphasis). For the latter, his RDU was partly motivated by normative considerations derived from the experimental results discussed below (see Quiggin 1982, p.325), an approach that, as he later lamented, "unfortunately has not been adopted by subsequent experimenters" (Quiggin, 1993, p.129; it should be noted that in this passage, he comments on Slovic and Tversky's design, but mistakenly references MacCrimmon and Larsson).

2009; 2014).

The work of Allais did pave the way for the "the experimental translation of rationality" as Mongin (2014) puts it. In these experiments, the decision makers are not decision theorists anymore, but ordinary subjects (or business executives in the pioneering one, MacCrimmon 1968). They confront both Allais' decision problems (along with others) *and* the arguments of decision theorists debating the rationality of the independence axiom. I wish to highlight three points about these experiments that are related to the methodological perspective of this dissertation.

The first point concerns Kenneth MacCrimmon and Stig Larsson's (1979) experiments. In their design, After his choices were made, the decision maker rates his degrees of agreement with twenty decision rules, which are ordinary language translations of either the formal axiom usually violated (e.g., the independence axioms), or their implications, or alternative decision rules in contradiction with EU (e.g., from Allais' theory on the specific role of Certainty, and non-independence of Probability and utility). Notice the communicative structure, whereby the issue of translation from decision theorists' formal language to the decision makers' ordinary language becomes crucial .

The second point concerns MacCrimmon's (1968, sect. III) pioneering experiment. In his design, After his choices were made, the decision maker is presented with the descriptions of two choices that are said to have been made by Other People, along with the respective explanations of them reflecting Allais' arguments on the specificity of Certainty or reflecting the independence axiom. The decision maker is asked to select the choice and explanation he prefers (and to try to criticize both): this step tends to make more people go for Allais, even those who respected the independence axiom. Then, in a last step, MacCrimmon conducts an interview with the decision maker to make sure he understood the task, and to know more about the reasons for his choices and eventual switch of choices. This sort of explicit communicative structure between the decision maker and the decision modeler seems very much in line with the perspective defended in this dissertation. But it rather points the limits of this perspective, because the results are rather puzzling: though this was not intended, decision makers decided to re-reverse their choices during the interview, to side against Allais. This shows (as most experimenters know well) how natural conversation, though not everyday ordinary language *per*

se, does not fit with the type of behaviors the decision modeler seeks to observe in the lab.

The third point is about the work of Tversky in this literature; it concerns how Slovic and Tversky's (1974) experiments provide a way out of the problems in the previous point. In one experiment, AFTER his choices were made, the decision maker is introduced counter arguments to his choices with the following instructions: "[n]ow that you have made your choices you might be interested in reading what a prominent decision theorist has to say about this problem" (pp.369-70). If they violated the independence axiom, they were introduced to a careful explanation and defense of this axiom, narrated in the first-person by "Dr. S.", reflecting Savage's arguments. If they did not violate the independence axiom, they were introduced to "Dr. A."'s careful explanation of why violating the axiom would have been rational, reflecting Allais' position. Then they had to choose again. Though there was much more people going for Allais in the first choices, the argumentative interlude did not change this so much. In another experiment, decision makers are presented with both argumentations with a careful step-by-step explanations of the arguments by the experimenter, which they had to rate AFTER they make their choices. On the overall, Allais' position was rated higher (more convincing), yet people (include those rating Allais' position high) chose more like Savage this time (MacCrimmon and Larsson's [1979, Fig.4] results display the same curious tendency).[27]

Despite the messiness of these results, they show at least that the traditional defense of the independence axiom fails, i.e., it is not the case that 'if people understand it then they tend to accept it, and then, with careful reflection, not to reverse their preferences'. The independence axiom is not always violated, yet its normative appeal to ordinary people, or lack thereof, remains an open question. Recently, a careful dissection of these experiments has been done by Mongin (2014), who offers a plea for their re-introduction into experimental economists' lab. Following Mongin (2014) I would like to propose potential experimental designs that may be worth implementing. In this vein, notice one feature of Slovic and Tversky's (1974) experiment: their willingness to present the matter with a specific type of external validity, namely to reflect the actual debate that is being carried out in the scientific community. This is especially vivid

[27]This tendency is further discussed in a comment on Mongin (2014), see Jullien (2016b). Also, though non-probabilistic uncertainty lies outside the scope of this dissertation, it should be noted that in the version of the problem with unknown probability (in the distale zone of the POSSIBLE), people overwhelmingly rated Allais' position higher, and chose consistently with it.

in the second experiment, where the two argumentations presented at the same time are based on the controversial conversations that are actually taking place outside of the lab. The decision makers' reactions and reasoning are potentially having an impact on (at least a part of) this debate (e.g., if there would have been 100% agreement on one side, the debate might have been settled). But this is not how the decision problem is presented to them. Presenting the decision problem as such, that is, as underlying actions constitutive of a scientific debate, might yield interesting outcomes, and would be in line with economists' experimental tradition of not deceiving decision makers regarding the purpose of their tasks. In short, the proposal is to present the tasks for what they are, i.e., as part of a controversial inquiry into the meaning(s) of economic rationality, and into the potential correspondence between economic agents and economic theorists on the issue.[28]

One obvious objection to this proposal – maybe even against the whole project of reviving this type of experiments – is that it is plagued with 'ambiguities' given the elusive and controversial aspects of the main measurement targets, namely the subjects' reasons, or reasoning; or worse, the meaning(s) of economic rationality. But as Sen puts it in his discussion of the debate around EU, "the notion of rationality involves inherent ambiguities [...] with or without uncertainty", and "to make these ambiguities clear" is worth pursuing (2002, chap.6, p.230). From Sen's

[28]To bring some concreteness to this proposal, let me relate how, with Dino Borie (the co-author of the last chapter of this dissertation), we took a rough shot at implementing it with our students. On a few occasions during a semester, we conducted choice experiments on framing effects. Besides telling students that there were no wrong answers, that it was for our own research but had nothing to do with the course, we also told them that if they were interested we would explain to them what this was all about at the end of semester. So at the end of the semester, we explained them what this was all about. Roughly, we presented several pairs of choice from the literature, asking them how they would choose and if they remembered how they had chosen, and telling them how most people choose. All along, we explained that there were never 'good' answers, but that some theories and theorists thought they were more or less 'rational' patterns of answers, how they disagreed, etc. This was, in a way, an exercise in the popularization of science (which, in passing, is interestingly argued to be of more interest for the scientist than for the reader by Jurdant 1973). At the end, we simply asked them again: 'Now that you know all that, does it make a difference to the way you see the problem? Would you change your original choices?'. The answers are not very representative because this was conducted non-rigorously (more for our own curiosity, and to actually test some weird axiom of our model, than for the sake of an eventual publication). Roughly, most of them stuck with their initial choices and a few had interesing rationale for their choices that I never thought about before despite having read the philosophical literature on the topic. It does not seem impossible to introduce controls for the obivously potential self-confirmatory biases in such procedures (viz. MacCrimmon's 1968 problems with the interview). The point here is that the proposition made in the paragraph is feasible. Furthermore, it can be given an epistemological justification through the theory of reflective equilibrium and the rationality debate in cognitive sciences (which is, as in economics, also around Kahneman and Tversky's work, among others); this is an argument I am currently developing in a working paper. Finally, it can be argued that the proposal made here is in line with two recent trends related to behavioral economics. The first one is the development of 'experimental philosophy' (Nagatsu 2013 provides an introduction directed to economists). The second one is with the collective decision processes in "deliberative democracy" discussed by Lowesntein and Ubel (2008, p.1807), where representative citizens are making decisions informed by scientific results on decision making biases; my proposal is a variant at the individual level, with information about the debates around the results.

reflections and given the experiments discussed above, there are at least two ambiguities that may be clarified with some specific experimental conditions. First, how much acquaintance with formal language is to be required in tracing the rational/irrational line (ibid)? Mongin (2014) contrasts Allais' exogenous selection of subjects (only highly acquainted subjects participate to the experiments) with MacCrimmon and Larsson's nearly endogenous selection of subjects (using within-subject data that seek to measure the coherence between choices and approval of the underlying rules and reasoning of choices, the more coherent the more rational). The recent focus on inter-individual differences and intelligence in the literature around behavioral economics has already implemented strategies that mix exogenous and endogenous selection of rational subjects (a brief discussion of which is made §4.2.1 in chapter 4 below). Second, "if anyone does claim that Allais' reasoning [...] are "erroneous," he has to show why the apparent justification is not "really" acceptable" (Sen 2002, chap.6, p.232). In other words, beyond presenting subjects with Allais' or Savage's reasoning, one could also present them with what is wrong with either reasoning from the other's perspective. In short, beyond arguments for or against a given axiom (as in Slovic and Tversky 1974), the decision makers should also be provided with the existing counter-arguments to them (which is slightly different from being asked to produce such counter-arguments, as in MacCrimmon's 1968 experiments).[29]

Turning to time preferences, I would like to argue that there is an internal requirement of reflective reasoning in dynamic consistency, by contrast with how this requirement has been externally imposed on the independence axiom by economists as an evaluative criterion. This is especially clear in the pioneering work of Strotz who introduced the notions of time consistency and inconsistency used in the previous subsections. The main result for which Strotz is usually credited is that *only* the theoretical convention of exponential discounting (and no other functional form) ensures dynamic consistency (see Lapied and Renault 2012, sects.1-4, for some subtle nuances to this observation). This result is indeed one of the main sources of value judgments defending the normative soundness of exponential discounting. Strotz's original discussion (1) gives some depth to this proposition (mainly because of the way he entangles facts and values) and at the same time (2) suggests a classical argument against being consistent with

[29]There is a third ambiguity which is too deep to be discussed in this dissertation, regarding the uses of counterfactual reasoning in the EU debates (see Sen 2002, chap.6, p.230; Bradley and Stefansson 2016).

oneself through time (namely, that there is no '*one*self' through time, but a series of selves). Let's discuss these two points in turn.

Firstly, Strotz argues from *p*sychology that we are not naturally dynamically consistent in our intertemporal choices. But we seem to value dynamic consistency because there are many empirical examples of behaviors where the decision maker *tries* to avoid dynamic inconsistency. In short, the theoretical convention of dynamic consistency is justified by value judgments of rationality and irrationality from the factually observable behaviors of decision makers who try but not necessarily succeed to be dynamically consistent. Strotz classifies such behaviors in two categories: *precommitment* and *consistent planning* (1956, p.173). The former includes cases such as 'Christmas Clubs' where the decision maker save some money with overly restrictive conditions in terms of interest rates and withdrawal fees so as to be sure to have enough money to buy Christmas present in due time. The central point about precommitment is that decision makers precommit to a plan to make it harder to revise it in the future *because* they anticipate the unpleasant future consequences of the plan will trigger later revisions leading to dynamic inconsistency. By contrast, consistent planning are cases when the decision maker anticipates his inconsistency in future revision of the plan and chooses another initial plan accordingly. Strotz offers no concrete examples of consistent planning; he only tries to offer a formal solution to it by treating the decision problem backwards. Notice that, on Strotz's accounts of both precommitment and consistent planning, dynamic consistency is the fruit of the decision maker's reflective reasoning about how he *would have* behaved in the FUTURE without having this very reflection, which is NOW changing how he will *actually* behave in the FUTURE. Strotz's analytical result that only exponential discounting ensures dynamic consistency, added to the entanglement of facts and values in his characterization of dynamic consistency, provides a non-trivial normative justification for the theoretical conventions of exponential discounting and their implications.[30]

[30]In a contemporary multiple selves' terminology, if the evaluating self t can undertake a precommitment action that retrains the actions of the future selves, then this game formalizes Strotz's problem of precommitment, otherwise it formalizes consistent planning. Though the language of selves was not used strictly speaking, the "sophisticate"/"naïve" terminology introduced in the previous subsection indeed stems from (Pollak 1968) the literature generated by Strotz's consistent planning problem, which, like Laibson, focus only on the sophisticated case (see esp. Peleg and Yaari 1973 and Goldman 1980). O'Donoghue and Rabin's (1999a; b; 2001; 2003; 2005; 2006) contributions can be seen as systematically contrasting the sophisticated case with the naïve cases, and thus pointing the implications of the latter case that has been under-examined in the consistent planning literature. One non-trivial implication is that there is never consistent planning (by definition) and it is also harder to

Secondly, Strotz raises the point that "[t]o-day it will be rational for a man to jettison his "optimal" plan of yesterday, not because his tastes have changed in any unexpected way nor because his knowledge of the future is different, but because to-day he is a different person with a new discount function – the old one shifted forward in time" (p.173). The main normative implication for an economist being "[t]he interpersonal aspect of the intertemporal problem" (p.179), namely that:

> "consumer sovereignty has no meaning in the context of the dynamic decision-making problem. The individual over time is an infinity of individuals, and the familiar problems of interpersonal utility comparisons are there to plague us. " (ibid)

In other words, Strotz's position is that *if* the decision maker is not the unit of agency through time, then dynamic consistency *loses normative* appeal, and by implication of Strotz's main *analytical* result, so does exponential discounting; though the *facts* remain that decision makers often seek to achieve dynamic consistency, which suggest that they are the unit of agency through time after all..

The relation with the 'primacy of time preferences' in the work of Kahneman is this. The experimental literature on experienced utility is taken to show that, partly because they fail to correctly remember NOW what they have experienced in THE PAST, decision makers usually fail to anticipate NOW what they will experience in THE FUTURE. Since this is a prerequisite to anticipate how they will behave then, dynamic inconsistency is more pervasive and inevitable than Strotz thought. And dynamic inconsistency is not restricted to THE FUTURE, but to consequences that occurs just a little bit LATER after the choice. Thus, the systematic biases in anticipation, formalized in Loewenstein, O'Donoghue and Rabin's (2003) model of "*projection bias*", also contend to explain inconsistency in decisions that were usually explained in atemporal frameworks, such as endowment effects (e.g., one fail to anticipate the effect of loss aversion in his evaluation, see Loewenstein et al. 2003, p.1214).

On its normative dimension, the primacy of time preferences in the work of Kahneman has been criticized on the basis that the maximization of instantaneous hedonic experiences does not capture aspects of decision making that are normatively meaningful to decision makers. Among these aspects, is the normative value of narrativity. That is, the value of planning one's life

observe directly since naïves don't undertake precommitment (as they think they are more time-consistent than they actually are, see esp. O'Donoghue and Rabin's 1999a, sect. V).

as a story to be narrated which involves (among other things) unplanned drama that we may or may not overcome (see the references in Loewenstein and Ubel 2008, p.1803). The upshot of the argument is that dynamic inconsistencies can be justified as rational by a consistent story. More precisely, one can construct a consistent story to justify dynamic inconsistencies by arguing against *the necessity* of dynamic consistency (either in its behavioral or 'experienced' interpretations) as a component of the good life. Notice that this does not imply that *all* dynamic inconsistencies are necessarily rational, just that *some* are, hence dynamic consistency is not *necessarily* rational. To illustrate, imagine that I choose a plan in which I work as much as possible on my dissertation during the last month before giving the final manuscript back over other plans where I sometimes party with my friends at night (and do not work the day after) or where I go on some road-trips to Italy and do not work during them. For this given situation, consider the following two violations, which *ceteris paribus*, imply the same objective outcome (e.g., minus one day of work on the dissertation): partying with my friends one night *versus* an impulsively unplanned road-trip of one day to Italy with the same friends. It could be argued that any story cannot make the place of the former violation in my life important enough (without 'lying to myself') to judge the underlying behavior as rational, while the reverse is true for the latter violation. The point is that this criticism raised against the primacy of time preference in the work of Kahneman extends beyond it, i.e., to dynamic consistency in general. Developing this argument at length would take too much space as it involves situating behavioral economics within broader contemporary debates (notably in the philosophy of the social and cognitive sciences, see Strawson 2004) about the issues underlying intertemporal choices.[31]

Turning to the 'primacy of social preferences' in the work of Sen, consider the following question which is often raised in discussions of Sen's commitment (see Peter and Schmid 2007). If commitment is a species of reason, then what are the specific characteristic of that species

[31]Notice that for the criticism of this paragraph to hold, one must restrict notions such as 'consequences', 'outcomes' and the like to objective ones in the sense that their descriptions cannot contain references to mental states (this issue is discussed in the conclusion of chapter 4 and in chapter 5). Notice also that John Davis (2009a; 2011) and Don Ross (2005; 2014) have undertaken to place behavioral economics within the debates on narrativity and identiy in the philosophy of cognitive and social sciences. We critically assess their work in a working paper co-written with Tom Juille. It should finally be noted that, at least implicitly and mostly by drawing on the "*narrative fallacy*" developed by Nassim Taleb in his popular book *The Black Swan* (2007), Kahneman (2011, pp.199-200, 218-221, and chap.36) took a stance on this narrative criticism. He agrees that narrativity is indeed an important source of motivation for people's behavior, so it should be explained positively. But he argues that it does not in fact tend to contribute to make them happier and has other negative consequences, and so it should not be used as a normative criterion.

from the other ones (self-centeredness and sympathy), especially regarding the way it motivates behavior? One part of the answer (but by all mean not the full answer) is that rational commitments can be the fruit of *social identity* regimented by reasoned scrutiny. Expressed in his words, rational commitments are possible through "a sense of "identity" generated in a community (without leading to a congruence of goals)", and managing the non-congruence with others' goals is achieved through "the fourth aspect of the self" besides the three requirements of self interest and not imposed in standard models, i.e., "one's own reasoning and self-scrutiny" (Sen 2002, chap.5, p.219 for the first quote; then chap.1, p.36; see also Davis 2007a).

Sen defends a plural conception of identity where what makes the decision maker the unit of agency is his reasoned scrutiny about all the identities he has from belonging to different social groups. It is however worth keeping in mind that this does not exhaust the specificity of rational commitments because "identity-based reasoning, even of the most permissive kind, including the identity of belonging to the group of all human beings, must, however, be distinguished from those arguments for concern that make no use of any particular *shared membership*, but nevertheless invoke ethical norms (of, say, kindness, fairness or humaneness) that may be expected to guide the behaviour of any human being" (Sen 2009 fn pp.142-3; see also 2006, pp.22-3).

Social identity underlies a methodological issue about the unity of agency which makes a connection with time preferences, as we saw a similar issue arise there as well. Put simply, while the issue for time preferences is 'what makes it the case that I will be Me in The Future (and was Me in The Past)', for social preferences it is 'what makes it the case that I am Now Me and not You, and that He is (or They are) not with Us?'. The precommitment behaviors discussed by Strotz and the commitment behaviors related to social identity discussed by Sen are only a tiny part of the answers to these two questions.[32]

To conclude this subsection, it can be argued that there are characteristics of reasoned scrutiny involved in the experimental translation of rationality, in both dynamic consistency and its violations rationalized by narrativity and in commitments, even those that the decision maker's derived from his identity. Several complementarities suggest themselves. For instance,

[32]They two questions have been formulated in a more sophisticated way and deeply scrutinized by John Davis in many publications (systematized in his 2003 and 2011 books; see esp. 2007a). His contribution is in line with this chapter for two reasons. First, Davis takes the two questions *together*. Second, he argues for a consistent answers to *both* questions by building on Sen's work on capability and commitment (with a specific attention to reasoned scrutiny that Davis develops in his account of 'reflexivity').

the experimental translation of rationality can be applied in experiments on time and social preferences; or the type of dialogue between Dr. A. and Dr. S. in Slovic and Tversky's experiments can be used as inspiration to write experimental instructions in a narrative manner that speaks more directly to decision makers than either open ended questions or neutral ones (for the purpose of checking reasons for choices at least). The perspective taken in this subsection suggests that an articulation of the notions of identity and narrativity can provide an helpful critical background to assess the normative dimensions of economic behaviors under uncertainty and especially over time and regarding other people (for developments in that direction, see Ross 2005; 2014; Davis 2009a; 2011). Finally, note that this perspective is somewhat 'external' in the sense of using arguments developed by authors who do not participate to the contemporary standard *versus* behavioral economics debates – though it is not external in the sense of *not* proposing alternatives alien to these debates, quite the contrary.

Conclusion

This section has provided an empirical and theoretical picture of the main challenges posed by behavioral economics to the standard accounts of behaviors under uncertainty, over time and regarding other people. Despite being three different areas of research, we saw non-trivial but implicit connections across the three dimensions. Both in terms of the issue of interdisciplinarity between economics and *P*sychology and of the positive/normative issues, the dimensions of uncertainty and time have more similitude than uncertainty and other people or time and other people (recall that we focus on the challenges posed by behavioral economics, not on standard economics taken as a whole). Notably, while under uncertainty and over time there are two well identified sources of inspirations from *P*sychology (respectively Kahneman and Tversky, and Ainslie) and a rather strong adherence to the doctrine of consequentialism from a normative perspective, both these aspects are less present regarding other people (but they are there). Throughout, we have emphasized by means of Rastier's typographical convention that one condition of possibility for the discussions provided here to be intelligible in the first place is the marking of linguistic distinctions within and among his table presented in the introduction of the chapter. Finally, it can be said that the critical exercise conducted in the last subsection was 'external' in the sense that the arguments developed here are not derived from authors

that participate to the contemporary standard *versus* behavioral economics debates. The next section can be seen, by contrast, as continuing this critical exercise 'internally', i.e., by drawing on contemporary authors participating to these debates.

2.2 Interactions across the three dimensions

Actually, the perspective taken in this section can deemed 'internal' in two senses. There is the sense just mentioned, i.e., putting together a number of contributions from within the standard *versus* behavioral debates. But it is also internal in the sense that what most of these contributions do is to 'internally' fold the three dimensions over each others, by contrast with putting them in perspective with a fourth 'external' entity, e.g., the *p*-&-*P*sychological constructs of motivation, attention or emotion (see e.g., Bruno 2013) – or even narrativity and identity as in the previous subsection. This section characterizes some details of the post mid-2000s emergence of interests in interactions across the three dimensions. Because many of the contributions discussed here do not discuss each others, the main goal of this subsection is to state the issues and argue that they are non-trivial for the behavioral *versus* standard economics debates; the goal is *not* to discuss a possible general solution to these issues (this will be the object of the next chapter). A picture of the main empirical regularities characterizing the post mid-2000s' interactions across dimensions movement is first presented (2.2.1). Then, in order to understand what the theoretical and normative implications of this picture are and are not, it is put into perspective with three sets of contributions. The first one is on *separability* as the main theoretical convention and source of normativity underlying interactions across dimensions within standard models (2.2.2). The second one is on early theoretical and normative challenges to separability – from the *timing of uncertainty resolution* and *conflicts between dynamic consistency and consequentialism* – which have recently gained interests around behavioral economics because of the post mid-2000s' interactions across dimensions (2.2.3). The last one is on more recent (quantitative) measurement and (qualitative) theoretical issues that relates the picture with both the issue of *interdisciplinarity* and an explicit version of the issue of the *primacy* of one type of preference over the others (2.2.4).

2.2.1 Altogether and naked again: the main empirical regularities

In the same fashion as what was done in the first subsection of the previous section, this subsection tries to present a structured picture of the main empirical regularities about interactions across the three dimensions. Just as one of the specificity of the picture there was to discuss the three dimensions altogether instead of separately (as is usually the case), the picture here discuss the three pairs of interactions (i.e., time-risk, risk-social, social-time) altogether instead of separately (as is usually the case). It is also naked in the sense that theoretical and methodological considerations have been kept at minimum for the same reasons evoked there. And again, the examples have been carefully selected for the way they make the 'logic' of the regularities transparent. The main goal of this subsection is to flesh one feature of this logic, namely that there are directions of introduction of one dimension in another one, due to historical reasons pertaining to the empirical structure of the classical challenges discussed in the previous section.

The example of a decision problem involving time and risk in the introduction of this chapter was taken from economists Manuel Baucells and Franz Heukamp (2010, table 1). It can be said of this example that it *introduced time in the dimension of uncertainty* because the structure of the decision problem is the one traditionally used to study common ratio effects, which played a crucial role in the literature around risk preferences. Indeed, Baucells and Heukamp first replicated the common ratio effect ($[A \succ B]\&[D \succ C]$) from the previous section with the dimension of time made explicit and the following monetary values (where the common ratio dividing the probabilities is 10):

A: The Certainty of winning €9 now
B: 80% chance of winning €12 now
C: 10% chance of winning €9 now
D: 8% chance of winning €12 now

Then they show how varying the dimension of time can, with respect to the standard model (EU), restore consistency in risk preferences ($[B' \succ A']\&[D' \succ C']$), as in the following decision problems where all the consequences from the previous problems are delayed in The Future:

A': The certainty of winning €9 in 3 months
B′: 80% chance of winning €12 in 3 months

C': 10% chance of winning €9 in 3 months
D$'$: 8% chance of winning €12 in 3 months

One of the main inspirations of Baucells and Heukamp's experiment was a series of experimental results obtained fifteen years earlier by psychologists Gideon Keren and Peter Roelofsma (1995). The latter paper is indeed one of the main trigger (with ten years of delay) of the mid-2000s emergence of interests in the interactions between time and risk preferences around behavioral economics. Keren and Roelofsma (1995, exp.1) also *introduced risk in the dimension of time*, through the decision problem used to study the common difference effect discussed in the previous section. They first replicated the effect ($[A \succ B]\&[D \succ C]$) with the dimension of risk made explicit and the following monetary values (where the difference in time is 4 WEEKS):

A: The certainty of winning €100 now
B: The certainty of winning €110 in 4 weeks
C: The certainty of winning €100 in 26 weeks
D: The certainty of winning €110 in 30 weeks

Then they show how varying the dimension of risk can, with respect to the standard model (EDU), restore consistency in time preferences ($[B' \succ C']\&[D' \succ C']$), as in the following decision problems where all the consequences from the previous problems can be won with PROBABILITY .5:

A': 50% chance of winning €100 now
B$'$: 50% chance of winning €110 in 4 weeks
C': 50% chance of winning €100 in 26 weeks
D$'$: 50% chance of winning €110 in 30 weeks

We have seen that the interactions between risk and time preferences can go both ways: time preferences can restore consistency in risk preferences (Baucells and Heukamp 2010; Keren and Roelofsma 1995, exp.2), and risk preferences can restore consistency in time preferences (Keren and Roelofsma 1995, exp.1). These interactions are indeed often discussed altogether in the post mid-2000s literature on the interactions across dimensions, focusing mainly on the effects just presented (though there are others that will not be discussed here, see the references in the next subsection). But, on the one hand, they are discussed apart from interactions involving

the dimension of other people, and, on the other hand, the different interactions involving the dimension of other people are often discussed separately from one another. We now turn to the latter, starting with how risk and time have been introduced in the dimension of other people.

An instance of the *introduction of risk in the dimension of other people* is the dictator games of Michal Krawczyk and Fabrice Le Lec (2010). They construct a "probabilistic dictator game" (p.500) where a decision maker has to allocate 10 tokens between himself and the recipient in different conditions. In a deterministic condition mimicking standard dictator game under certainty, one token is worth 10% of the consequences of a predetermined allocation, e.g., if the predetermined allocation is €20 for the decision maker and €20 for the other, then keeping the 10 tokens imply €20 for the decision maker (100% of his consequences) and nothing for the other (0% of his consequences), while giving 5 tokens to the others imply €10 for the decision maker (50% of his consequences) and €10 for the others (50% of his consequences). In another condition with risk, one token is worth 10% chance of winning the consequences of a predetermined allocation, e.g., (with the same €20/€20 allocation) keeping 10 tokens imply the certainty (100% chance) of winning €20 for the decision maker and the impossibility (0% chance) of winning €20, while giving 5 tokens to the others imply the lottery with 50% chance of winning €20 and 50% chance of winning nothing for the decision maker, and the same lottery for the other. The decision makers who gave at least one token in the first case (they gave 2.43 on average) gave significantly less token in the second case (they gave 1.98 on average). With other results from other conditions, the tendency is that the greater the uncertainty introduced the closer to self-centeredness the decision makers tend.

An instance of the *introduction of time in the dimension of other people* is in the dictator games of Jaromir Kovarik (2009). He constructs dictator games where the decision maker has to allocate €6 between himself and the recipient in different conditions. In an immediate condition mimicking the standard dictator, the decision maker and the other both receive their money at the end of the experiment. The other conditions vary the delay with which the money is received, from two days to twenty two days. The tendency is that the greater the delay introduced, the closer to self-centeredness the decision makers tend, with virtually everybody being strictly self-centered after fourteen days of delay (i.e., all give €0 from fourteen till twenty two days of delay).

Again, other effects of risk and time in the dimension of other people have been shown with other games but they will not be discussed here (see the references in the next subsection). The point is that introducing risk and time in the dimension of other people can push decision makers to abide by the requirement of self-interest that was initially violated in the dictator game, i.e., self-centeredness.

We can finally turn to how other people can be introduced in the dimensions of risk and time. An instance of the *introduction of other people in the dimension of risk* is in the experiments of Gary Bolton and Axel Ockenfels (2010), with decision problems such the following two pairs (A *versus* B and A' *versus* B'):

> **A**: the certainty of winning €9 for Me
> B: 50% chance of winning €16 and 50% chance of winning nothing for Me
> A': the certainty of winning €9 for Me and €16 for Him
> **B**$'$: 50% chance of winning €16 for Me and €16 for Him
> *and* 50% chance of winning nothing for Me and nothing for Him

Notice how the modal risk preference for oneself ($A \succ B$) reverses when the other is introduced ($B' \succ A'$) even though the consequences for oneself are invariant across the two pairs. The second pair just adds inequality in the sure object of choice and equality in the risky one. Thus it seems that a preference for equality (or against inequality) interacts with risk preferences, creating inconsistency in the latter (at least with respect to the standard model). So by contrast with the two first examples of this subsection, the interaction here creates a further inconsistency rather than restoring consistency.

To my knowledge, *there is no experiments introducing other people in the dimension of time*, at least in the same fashion as the experiments discussed so far in this subsection, i.e., where the introduction of other people in an intertemporal choice would create a conflict that would have an effect on time preferences. This is surprising for two reasons, which I will briefly discuss and for which I will propose decision problems that would easily be implemented in lab experiments. First, such decision problems are not hard to imagine. For instance, one instance drawing on the previous choice experiment would look like this:

A: €9 for Me now
B: €16 for Me in a week
A': €9 for Me now and €16 for Him now
B': €16 for Me in a week and €16 for Him in a week

The second reason why the absence of such experiments is surprising is because the relation between discounting and altruism (cf. the problems of intergenerational altruism mentioned in the previous section) is currently gaining great theoretical and empirical attention regarding climate change, and this has directed the attention of macroeconomists interested in this issue of so-called 'social discounting' to look at the behavioral economics literature on individual time preferences (see the references in Peroco and Nijkamp 2009). An important problem related to individual time preferences in social discounting is illustrated by Sen's (1961, pp.487-9) "*Isolation Paradox*", which could also be of interest to behavioral economists. The original thought experiment runs roughly as follows. A decision maker who will be dead in twenty years prefers one unit of consumption now to three units of consumption in twenty years, even though he cares about the future generations, i.e., three for the other is not enough for him to sacrifice one for himself. Another individual confronted with the same problem comes to the decision maker and says: 'if you save your three units, then I will do the same, if you consume your unit now, then I will also do the same'. The decision maker changes his preferences, and Sen argues that this reversal is perfectly rational, even without deep moral considerations, i.e., only on quantitative grounds, sacrificing one unit for himself for six units to the other is enough. This would easily translate in the following individual decision problem:[33]

A: €10 for Me now
B: €30 for Him in a week
A': €10 for Me now and €10 for You now
B$'$: €60 for Him in a week

Summing up, we have seen in this subsection that interactions across dimensions can (with respect to the standard models) either restore consistency in risk, time or social preferences, or

[33]In psychology experiments the expression 'social discounting' is also used in the literature on 'social-distance' mentioned in the last footnote of the first subsection (e.g., Jones and Rachlin 2008). The example suggested here is somewhat close to the ones presented there (see esp. Yi et al. 2011), with the difference that the psychologists introduce more experimental manipulations aimed at measuring the social distance between Me You and Him (they asked to imagine who and how you would rank people on a list from 1 to 100).

produce further inconsistencies. Experiments have investigated interactions across dimensions by pairs, with an absence of interest in the time-social pair, at least in the same experimental fashion as the other ones have been studied. Although the effects presented here are not exhaustive, it seems that there has not been similar studies on interactions across dimensions in magnitude effects (though see the pattern of preferences called '*subendurance*' in Baucells and Heukamp 2012, table 1, lines 7-8) or in sign effects. The same hold for 'social distance' as defined in experimental economics (see the previous section), though psychologists have done such studies regarding the social-time pair with a somewhat different understanding (or at least experimental implementation) of 'social distance' (see the previous footnote). But the picture presented here is (I hope) enough to get a concrete taste of what is meant here by interactions across dimensions compared with the interactions within dimensions from the previous section, which were constitutive of the making of behavioral economics. The main contrast is that here the patterns of preferences are, taken altogether, in contradictions with *both* the standard model *and the behavioral alternatives* presented in the previous section.

2.2.2 Separability and the three dimensions altogether

In standard models, the implicit connections between the three dimensions are due to mathematical results and theoretical contributions of John Harsanyi (1955), William Gorman (1968) and Debreu (1960). The goal of this subsection is first to explain these connections by following John Broome's (1991) book-length investigation of the theoretical and normative interpretations of these contributions. Broome provides one of the rare discussion of the implicit connections in the standard model of the three dimensions taken altogether; we shall then to put it in perspective with two more recent contributions that also tackle the three dimensions altogether.

Broom's account focuses on the role of *separability* in the construction of functional representation ($U(.)$) of preferences ($\succsim$). We shall first illustrate that the standard models studied above share a common requirement of separability, namely *additive separability*, before we illustrate Broome's account of the interactions across dimensions though what he calls *crosscutting separability* (see also Grant and van Zandt 2009). In the dimension of uncertainty, preferences are separable between mutually-exclusive states, e.g., in a coin flip lottery (X), there are only two mutually exclusive states, namely of head *or* tail coming up (1 *or* 2, respectively), the consequences in

each being therefore mutually exclusive as well (you get x_1 *or* x_2). Additive separability requires, not only that each of these consequences be subevaluated separately, but that they can be added, i.e., $U(x_1, x_2) = u_1(x_1) + u_2(x_2)$. The independence axiom necessary for EU is thus a separability condition that further requires the subutilities of consequences to be weighted linearly by their respective PROBABILITIES, i.e., $U(X) = .5u_1(x_1) + .5u_2(x_2)$, and assumptions of a unique subutility function are often made, i.e., $U(X) = .5u(x_1) + .5u(x_2)$. In the dimension of time, preferences are separable between mutually exclusive periods of a plan (X), e.g., NOW *or* LATER (1 *or* 2, respectively), with mutually exclusive consequences x_1 or x_2. What is sometimes called 'time separability' implies additive separability, i.e., $U(X) = u_1(x_1) + u_2(x_2)$. Time separability is part of the axioms necessary for EDU, other axioms implying the weighting of the subutilities of the consequences to be an exponential function of time, i.e., $U(X) = u_1(x_2) + \frac{u_2(x_2)}{1+\rho}$, together with the assumption of a unique subutility function, i.e., $U(X) = u(x_2) + \frac{u(x_2)}{1+\rho}$. In the dimension of other people, preferences are separable between the mutually the exclusive people that are the recipients allocation (X), e.g., ME and YOU (1 and 2, respectively) receives consequences x_1 and x_2 that are mutually exclusive in that $X = x_1 + x_2$. In modeling whoever evaluate an allocation (e.g., ME, YOU, or A SOCIAL OBSERVER), additive separability between people is often implied by axioms on preferences or assumed directly, i.e., $U(X) = u_1(x_1) + u_2(x_2)$, together with various and usually controversial assumptions or axioms about how to weight and/or compare the (sub)utilities of people (see Sen 1977a; or Harsanyi 1988 on these issues).

Crosscutting separability concerns preferences over consequences that are distributed across dimensions. To illustrate what crosscutting separability means and requires, as well as the analytical connection it gives rise to, we shall use Broome's type of example that are, as he puts it, "conveniently symmetrical" (1991, p.62) on two dimensions only. To do so, consider the following 'temporal allocation' or 'social plan':

> *A*: €10 now and €10 soon for me, and €10 now and €10 soon for you
> *B*: €5 now and €20 soon for me, and €20 now and €5 soon for you

This decision problem can be represented as in Table 2.2, where the subscripts separated by a comma indicate the coordinates of what Broome calls the "locations" (related to the 'zones' of a dimension in Rastier's terminology), the first one is the time coordinate, i.e., location 1 is

		Time			
		Now	Soon	Now	Soon
People	Me	€10 ($a_{1,1}$)	€10 ($a_{2,1}$)	€5 ($b_{1,1}$)	€20 ($b_{2,1}$)
	You	€10 ($a_{1,2}$)	€10 ($a_{2,2}$)	€20 ($b_{1,2}$)	€5 ($b_{2,2}$)
		A		B	

Table 2.2: A conveniently symmetrical example for crosscutting separability

Now and location 2 is Soon, the second one is the people coordinate, i.e., location 1 is Me and location 2 is You.

Crosscutting separability implies that you can evaluate the row and the column of this table separately. Because crosscutting separability implies additive separability, there is at least four subutility functions evaluating the consequences here, i.e., $u_{1,1}(.)$, $u_{2,1}(.)$, $u_{1,2}(.)$, $u_{2,2}(.)$. And the connections across dimensions occur because one evaluation on one dimension will depend on my preferences regarding the other dimension. This is the case because each consequence cuts across *two* dimensions (i.e., is localized in one location of each dimension), so that each consequence is evaluated by *one* subutility function, which captures both time and social preferences, which are thus identical. Let's illustrate this reasoning in more details.

Evaluating the time locations (i.e., the columns) separately means that I can subevaluate what happens Now separately from what happens Soon. When I subevaluate what happens now, I see that if I choose A, You and Me get the same thing (€10), and if I choose B, we don't get the same thing (I get €5 and you get €20). On this *time* location, my choice therefore depends on (at least one aspect of) my *social* preferences: what are my attitudes towards an *equal* treatment where both of us receive the same thing, and towards an *unequal* treatment where there is €20 for you and €5 for me? If I like equality, I will tend to choose A, if I don't mind about inequality (and/or if I genuinely care about other people at the expense of my self-centered self-interest) I will tend to choose B. When I sub-evaluate what happens Soon, I see that if I choose A, You and Me both get the same thing (€10), and if I choose B we don't get the same thing (I get €20 and you get €5). Again, on this time location my choice depends again on my social preferences: if I like equality I still tend to choose A, and if I don't mind about inequality (and/or if I genuinely care only about me), I will still tend to choose B.

Therefore, evaluated separately on the dimension of time, the result is the same whatever the time locations: if I like equality, I choose A, if I don't mind about inequality, I choose B. Thus, the result on the dimension of time depends on (at leas one aspect of) my social preferences.

Evaluating the people locations separately means that I can subevaluate what happens for ME separately from what happens for YOU. When I sub-evaluate what happens for me, I see that if I choose A I get €10 NOW and €10 SOON, and if I choose B, I get €5 NOW and €20 SOON. On this *personal* location, my choice therefore depends on (at least one aspect of) my *time* preferences *which does not correspond directly to discounting but to the elasticity of intertemporal substitution*: what are my attitudes towards an equally distributed flow of money between two points in time separated by a week, and towards an unequally distributed flow of a slightly bigger amount of money between the same points in time? If I like steady flows of money (where I get the same amount periodically), I will tend to choose A, but if I like unsteady flows of money (where I get more at some points though less at others), I will tend to choose B. When I sub-evaluate what happens for YOU, I see that if I choose A, you get €10 NOW and €10 SOON, and if I choose B, you get €20 NOW and €5 SOON. Again my choice on this location depends on (at least one aspect of) my time preferences: if I like equally distributed flow of money I will tend to choose A, but if I like unequally distributed flows of money I will tend to choose B. Therefore, evaluated separately on the dimension of people, the result is the same whatever the people locations: if I like equally distributed flow of money, I will tend to choose A, but if I like unequally distributed flows of money I will tend to choose B. Thus, the result on the dimension of people depends on (at least one aspect of) my time preferences. Notice that high discounting leads to a preference for plans with distribution of consequences skewed towards the present. By contrast, high elasticity of intertemporal substitution leads to a preference for plans with unequal distribution of consequences, not necessarily skewed, but usually assumed to be so towards the future because of positive interest (or growth) rates. Hence, from a revealed preference perspective, violations of dynamic consistency *can* come from a change in the former *and/or* the latter. Probably because the latter is meaningful in economics only with respect to further economic variables that are absent from experimental setups, measures derived from intertemporal experiments are interpreted as being about '*pure*' time preferences, i.e., discounting. Nevertheless, both parameters can non-controversially said

to constitute decision makers' attitudes over time. As Frederick, Loewenstein and O'Donoghue (2002, p.359) put it:

> "diminishing marginal utility (that the instantaneous utility function [...] is concave) and positive time preference (that the discount rate ρ is positive) [...] create opposing forces in intertemporal choice: diminishing marginal utility motivates a person to spread consumption over time, while positive time preference motivates a person to concentrate consumption in the present."

When we bring the evaluation on the dimension of time and the evaluation on the dimension of people together, we have a very simple and elegant conclusion. On the one hand, if I choose A, it reveals a preference for equality in the distribution of money across people, and a preference for equality in the distribution of money across time. On the other hand, if I choose B, it reveals a preference for inequality in the distribution of money across people, and a preference for inequality in the distribution of money across time. Hence disliking inequality in the distribution of money over time *and* across people "must be linked": "*[t]hey are really just two ways of describing the same pattern of preferences*" (Broome 1991, pp. 63-64, my emphasis). Adding one dimension to the previous reasoning generalizes it, the point being that, in standard models, *one* consequence that cuts across *two or three dimensions* is evaluated by *one* single subutility function that identify aspects of risk, time and social preferences with one another. The plausibility of the argument depends on crosscutting separability and "[i]f separability fails in one or the other dimension, the conclusion will fail too" (ibid). Of course, violations of the independence axiom, dynamic consistency and self-centeredness imply that separability tends to fail empirically in each dimensions. Broome is aware of this but his perspective is normative – and related to discussions of decisions about future unknown social positions under a 'veil of ignorance' where decision makers' risk and inequality aversion should be identical. Roughly, he argues from a version of consequentialism that violations of separability in the dimension of risk and other people are irrational; while those in the dimension of time are rationally justifiable because separability in that dimension carries a metaphysical implication of multiple selves which he finds unappealing.[34]

[34]Two further points are worth noting. Firstly, in a working paper, we have tried to illustrate how Broome's illustration of crosscutting separability reasoning carries over for the three dimensions altogether and/or for decision problems which are not "conveniently symmetrical"; it turns our that the elegance ans simplicity of the above reasoning is nontrivially reduced. Secondly, the main weakness that Broome acknowledges about crosscutting separability (in risk and people) is what he calls "[t]he rectangular field assumption" (p.80): that

Some empirical implications of the theoretical issues discussed by Broome have recently resurfaced in the normative economics of climate change, especially in a paper by Giles Atkinson, Simon Dietz, Jennifer Helgeson, Cameron Hepburn and Hakon Soelen (2009). One such implication is that standard social welfare functionals can be calibrated with data about either one of these three dimensions (i.e., risk aversion, elasticity of intertemporal substitution and inequality aversion), notably to calculate the social *discount rate* through which policies are evaluated. But in practice these are not identical because decision makers usually violate separability conditions. Hence calibrations with either one of the three types of data imply different policy recommendations from calibrations with either one of the two others. Thus, "[a]n important theoretical lacuna still exists because no model to date enables all three concepts to be disentangled simultaneously" (p.3). Furthermore, they argue against normative defense à la Broome for the identification of even two of the three aspects. Such a defense, on their account, presupposes conditions for reasoning that are unrealizable in practice (i.e., veil of ignorance-like). They further argue that normative justifications for the articulation (or lack thereof) among attitudes in the three dimensions need to be debatable in a democratic decision process, hence possibly going against consequentialism. Accordingly, they conduct a large scale internet survey to elicit these three attitudes of people regarding climate change issues, and found large difference and furthermore weak correlations among the three, contrary to "the standard welfare model [which] implicitly assumes a perfect correlation" (p.4). One implication is that an extra value judgment is needed in applied work, namely with data from which dimension should social welfare functionals be calibrated? This is obviously an issue where the stakes of the primacy of one type of preference over the two others looms large.

Finally, the paper by Güth, Vittoria Levati and Matteo Ploner (2008) quoted in the epigraph to this chapter tackles such correlations across the three dimensions using lab experiments. Decision makers are asked the minimum amount of money they would accept to sell each one of these objects of choice (i.e., their willingness to accept, WTA). Here are four examples of the sixteen objects of choices, where the lottery gives 50% chance of winning €16 and 50% chance

all the consequences need to be arranged in a cartesian product structure (rougly, as the above table), rendering some logically possible combinations of consequences empirically impossible (see Mongin and Pivato 2015 for a recent attempt at solving this issue). It could be argued from the perspective of this chapter that it does not matter if objects of choice are impossible or unreal *if* they remain describable and communicable to a decision maker (e.g., through a decision modeler belonging to an adverstisment company or a religion, for the better or worse of the decision maker's welfare).

of winning €38, 'later' is in three months, and 'you' is the decision maker (ibid, appendix, emphasizes and bolded terms are theirs):

> **You** get €27 for sure *now*, and **the other** gets the lottery *now*.
> **You** get €27 for sure *later*, and **the other** gets the lottery *later*.
> **You** get €27 for sure *now*, and **the other** gets the lottery *later*.
> **You** get €27 for sure *later*, and **the other** gets the lottery *now*.[35]

The main tendencies are that decision makers care about the other's consequences when their own consequences are immediate and sure, that their discount rates reveal much more impatience for their own consequences than for the other's, and that risk aversion usually goes with more impatience and risk seeking with more patience.

Notice that empirical studies traditionally investigate correlations between, on the one hand, either risk *or* time *or* social preferences, and, on the other hand, something 'external' to these preferences, e.g., some measures of economic success (as in Tanaka et al. 2010) or some measures of stability over the years in a given data panel of one type of preferences (as in Chuang and Schechtner 2015). By contrast, both Atkinson et al. (2009) and Güth et al. (2008) study 'internal' correlations among the three types of preferences. This approach is less traditional though not new *per se* (see the references cited by Jamison et al. 2012, p.14), but it can be argued that it is a growing trend motivated by behavioral economics (see esp. Dean and Ortoleva 2015). Two remarks about this trend are worth making, as they respectively introduce the objects of the next two subsections. The first one is that this trend is globally silent on the normative dimension (i.e., à la Broome) or implication (i.e., à la Atkinson et al.) of

[35] Here are the remaining twelve objects of choices:

> **You** get the lottery *now*, and **the other** gets €27 for sure *now*.
> **You** get the lottery *later*, and **the other** gets €27 for sure *later*.
> **You** get the lottery *now*, and **the other** gets €27 for sure *later*.
> **You** get the lottery *later*, and **the other** gets €27 for sure *now*.
> Both **you** and **the other** get €27 for sure *now*.
> Both **you** and **the other** get €27 for sure *later*.
> **You** get €27 for sure *now*, and **the other** gets €27 for sure *later*.
> **You** get €27 for sure *later*, and **the other** gets €27 for sure *now*.
> Both **you** and **the other** get the lottery *now*.
> Both **you** and **the other** get the lottery *later*.
> **You** get the lottery *now*, and **the other** gets the lottery *later*.
> **You** get the lottery *later*, and **the other** gets the lottery *now*.

Güth et al. are not interested in the absolute values of the WTAs, only in the differences between WTAs, thus the biases with this elicitation method do not matter much for their purpose.

their empirical results. The second one is that, though helpful to understand some underlying issues of the empirical picture presented in the previous subsection, quantitative correlations among dimensions are different from the qualitative issue of how new inconsistencies arise or disappear with the introduction of new dimensions into a given one. Notably, the former does not (at least) directly address how to get a parsimonious theoretical account of the patterns of preferences displayed in the empirical pictures of both this and the previous section taken together. Hence it could be argued that without such a qualitative work available, it is impossible to even represent the quantitative variations in risk time and/or social preferences underlying the choices of *one* decision maker abiding by the regularities of these two pictures, let alone investigate the correlations among them.

Summing up this subsection, the theoretical convention that decision makers attitudes in the three dimensions should be identical underlies standard models of individual behaviors. Value judgments from consequentialism can justify this theoretical convention and criticize its empirical violations, or *vice-versa*, value judgments against consequentialism can criticize the convention and justify the violations. Factual observations from correlational studies tend to show that, far from being identical, the three dimensions are not even strongly correlated. Finally, the marking of linguistic distinctions within and across Rastier's table is a condition of possibility for these empirical and theoretical discussions to be intelligible in the first place.

2.2.3 Timing of uncertainty and consequentialism *versus* dynamic consistency

The goal of this subsection is to briefly illustrate two classical challenges from within standard economics against the implications of separability discussed in the previous section, and to show that they have gained recent interests around behavioral economics.

The first one is the so-called issue of 'the timing of uncertainty resolution'. The classical paper stating the issue is David Kreps and Evan Porteus' (1978), who use the following intuitive example to illustrate it (p.185). A decision modeler proposes to flip a coin: if heads comes up you win €5 NOW and €10 LATER, if tail comes up you win €5 NOW and nothing LATER. Because there is a common consequence for now (€5), EU implies that the decision maker should be *indifferent* between flipping the coin NOW – so that the uncertainty about whether you get €10

or nothing later resolves now – or flipping it LATER – i.e., you get your €5 Now anyway but the uncertainty about whether you get €10 *or* nothing later resolves at the same time when you know which one you actually get. Kreps and Porteus' argument is twofold. On the one hand, they argue that there are good reasons for not being indifferent, notably a preference for early resolution can be justified from the ease it creates for budgeting NOW one's FUTURE expenditures. On the other hand, they propose an axiomatic framework from which models that analytically distinguish risk and time attitudes can be constructed. While such models have had their success in macroeconomics (see Gollier 2001, chap.20), they have not been discussed much around behavioral economics (though see Coble and Lusk 2010 and the reference therein for contributions in experimental economics).

The second challenge points some conflicts between consequentialism and dynamic consistency. Originally, the first instances of these challenges have been raised in normative economics (e.g., Diamond 1967; or more recently, Giraud and Renouard 2011; Bovens 2015; and more generally, see Mongin and d'Aspremont 1998). The structure of the argument has been applied to decision theory and Hammond's version of consequentialism in a classical paper by Mark Machina (1989). The following intuitive example illustrates such arguments (pp.1643-4). Mom has an indivisible treat, say one candy, she can give to *either* one of her children, Abigail *or* Benjamin. Mom is indifferent between (A) 'Abigail getting the candy FOR SURE' and (B) 'Benjamin getting the candy FOR SURE', and strongly prefers either one of these consequences to the one whereby (C) 'None of them gets the candy FOR SURE'. Mom however strictly prefers the PROBABILISTIC consequences – she can construct by flipping a coin – of ($D = (A, .5; B, .5)$) '50% CHANCE that Abigail gets the candy and 50% CHANCE that Benjamin gets the candy' to either one of the SURE consequences. Formally, Mom's preferences are $D \succ A \sim B \succ C$. So Mom flips a coin, it turns out that Abigail wins the candy, but Benjamin steps in and says: 'Mom, you told us EARLIER that you preferred 'to flip a coin' over 'Abigail getting the candy for sure' ($D \succ A$), NOW Abigail is getting the candy for sure so please respect you preferences and flip a coin'. In other words, Benjamin argues from (Hammond's) consequentialism that Mom should respect the independence axiom of EU by sticking to the part of her initial preferences that does not violate the requirement of linearity in the weighting of probabilities (i.e., preferring the gamble over *one* sure consequence, $D \succ A$). Mom answers "You had your chance!" and gives

the candy to Abigail. In other words, Mom argues that she is happy to violate consequentialism because she wants to be dynamically consistent with the part of her initial preferences that violate EU (i.e., preferring the gamble over the indifference between *both* sure consequences, $(A, .5; B.5) = D \succ A \sim B$). Machina (1989) argues at length that (Hammond's) consequentialism is a dynamic version of the very separability requirement that non-EU decision makers violate in the first place. Hence, if they have normative justifications for their violations, which is clearly the case for Mom here, another type of dynamic consistency than the one implied by Hammond's consequentialism is needed for non-EU decision makers. In Mom's example, the reason justifying her choice is a preference for fairness, indeed following the original literature in normative economics where these counter-examples to EU's normative dimension originally emerged. But Machina's point is more general. Roughly put, it is that 'history matters' in the following sense. Mom's preferences for 'giving the candy to Abigail *if* the uncertainty of 50% chance that Benjamin Benjamin gets the candy HAS BEEN BORNE (but not realized)' is already her preferences *ex ante*, i.e., before she knows the outcome of the coin flip. Hence by saying 'no' to Benjamin *ex post* she is being dynamically consistent with her *ex ante* preferences.[36]

I would like to make two points on Kreps and Porteus' and Machina's papers. The first concerns only Machina's paper. Reading Machina's paper through the lenses of interactions across the three dimensions, which is *not* the theme of his paper, one realizes how much interactions across the three dimensions are *sources of reasons* – or "*generators of reasons*" (Broome 1991, p.23, my emphasis) – in economics for both economic agents and economists. For economic agents faced with decisional conflicts in one dimension, e.g., uncertainty in Mom's case, another dimension is used to bring the justification that resolves the conflict, e.g., other people in Mom's case. For economists intending to formally represent such resolution in one dimension, contributions from other economists working in subfields that focus on other dimensions become relevant. In Machina's case, it is not only the contributions from normative economics in the dimension of other people, but also from intertemporal consumptions as he makes an analogy with the non-separable time preferences used there to ground his arguments (1989, pp.1644-1645).

[36] One classical argument against Machina's (made for example by Hammond) is that if the description of consequences is rich enough, including all the relevant psychological dimensions of the decision maker within them, then the independence axiom is not violated anymore. This argument is also made for Allais' paradoxes and other classical challenges from behavioral economics. We shall argue against this argument in chapters 4 and 5 on framing effects.

Interactions across the three dimensions are also sources of reasons in the sense that they makes intuitive decision theoretic arguments *appealing* in economics. Take for instance the well-known Dutch book (under uncertainty) or, more generally (under certainty) money pump arguments that are used to establish the *economic* rationality of not violating the independence axiom or the axiom of transitivity. The appeal of these arguments relies on ANOTHER (fictional) agent who has rather anti-social preferences toward the decision maker, as the former tries to put the latter to ruin OVER TIME (by repeated decisions).

The second point concerns the relation of both papers to the empirical picture presented in the first subsection of this section. Notice that the structure of the interactions across dimensions in both papers is not the same as the one in the empirical picture. Furthermore, both papers have been relatively disconnected from behavioral economics before the mid-2000s. However the theoretical outcomes of the empirical picture have made such a connection. This is so for Kreps and Porteus because Thomas Epper and Helga Fehr-Duda's (2015) theoretical unification of interactions across risk and time includes the former's phenomenon as one to be unified among the other ones constitutive of the empirical picture. And this is so for Machina because both Fudenberg and Levine's (2012a) and Stefan Trautmann and Wakker's (2010) theoretical contributions motivated by the empirical picture (respectively about the interactions across risk and other people, and across other people and time) make non-trivial uses of the former's reasoning. In a nutshell, Fudenberg and Levine show that all the models of non-self-centered motives discussed in the previous section cannot be extended under expected utility theory without violating a preference for *ex ante* fairness. That partly explains why none of these models can account for interactions across the dimensions of risk and other people presented in the first subsection of this section. Trautmann and Wakker make roughly the same point and emphasize how the phenomenon is better understood by connecting it with violations of dynamic consistency in the dimension of time. Finally, Andreoni et al. (2016) proposes a set of experiments to test whether decision makers tend to be consequentialist or dynamically consistent in social allocation problems under uncertainty. The answer is that it depends, among other factors, on whether they make their decisions *ex ante* (where they tend to be dynamically consistent) or *ex post* (where they tend to be consequentialist).[37]

[37] Rotemberg (2014, sect.6) provides a recent review of the post mid-2000s' interactions across the dimension

Summing up, it can be argued that Machina's arguments tried to establish even further the primacy of time over risk over social preferences in decision theory that was mentioned in the second subsection of the previous section. While Kreps and Porteus also went in that direction, it was merely by arguing against a full primacy of risk preferences. Though initially disconnected from behavioral economics, both these contributions are now being discussed around it because of the experimental results on interactions across dimensions. These facts, together with the value judgments against consequentialism and for fairness, are currently triggering some changes in the theoretical conventions constitutive of the study of individual decision making in economics. Again, a condition of possibility for the intelligibility of the whole process is the marking of linguistic distinctions within and across Rastier's table. Notice however that the interdisciplinary issue of the relations between economics and *P*sychology are somewhat absent from the contributions discussed so far in this section. In the next subsection, we shall discuss two sets of contributions in which not only this issue is explicit but also explicitly connected with the issue of the primacy of one type of preference over the others.

2.2.4 *p*-&-*P*sychology with the issue of primacy across the three dimensions

The goal of this subsection is to discuss two contributions, one on the qualitative issues underlying the interactions across dimensions and the other on the quantitative ones, in order to show how both the issue of interdisciplinarity between economics and *P*sychology and the issue of the primacy of one type of preferences over the others raise themselves in each case.

The contribution on the qualitative issues underlying the interactions across dimensions is Yoram Halevy's (2008) "Strotz Meets Allais", published in the *American Economic Review*. Halevy accounts for Keren and Roelofsma's (1995) results about the introduction of time in the dimension of risk and *vice-versa*. Halevy did not (just) weakened some axioms of EU or EDU to be compatible with the psychologists' results, his modeling strategy sought to be compatible with the *P*sychological explanations of these results by Keren and Roelofsma. The

of risk and social preferences. Trautmann and Vieider (2012) provide a very wide survey of the different ways by which other people can be introduced in the dimension of uncertainty in behavioral economics and social psychology. It should also be noted that Fudenberg and Levine's (2012a) contribution has motivated further empirical (see esp. Brock et al. 2013) and theoretical (see esp. Saito 2013) contributions on the interactions between risk and social preferences.

latter have to be understood against a background of claims defended in various publications by psychologist Howard Rachlin (a behaviorist whose work has non-trivially inspired Ainslie's). Part of Rachlin's work has already been referenced in some footnotes above on 'social distance' and 'social discounting', where he argued that decision maker's relations to other people are derivable from their relations to time, i.e., time is a primary notion explaining the modalities of interpersonal relations, esp. altruism. The claims against which Keren and Roelofsma argued were similar but about decision makers' relations to uncertainty, i.e., being derivable from their relations to time. Keren and Roelofsma argued that there might not be any primacy of one over the other, and that even if so, the other way around is more likely. The key argument that they want to defend in their paper is that THE FUTURE is *necessarily* UNCERTAIN for various reasons but at least because of *the omnipresent possibility of sudden death*, while UNCERTAINTY can also be in THE PRESENT which is thus *not necessarily* CERTAIN. Furthermore, the inherent uncertainty of consequences in the future stemming from the possibility of sudden death can only disappear through the IMMEDIATE experience of those consequences.

This argument plays a central role in Halevy's contribution, where the future is modeled "as a random process that has a positive probability of stopping at any given period" and "may be interpreted as the hazard of morality" (2008, p.1145). This captures the ("implicit") risk inherent in the future, which is absent in the present (i.e., equals one) so that the non-linear weighting function gives disproportionate weights to the present, inducing a present bias and dynamic inconsistency. The model nonetheless uses an exponential discount function, hence not the whole of time discounting stems from risk preferences. Thus, Halevy's model can account for the facts that quasi-hyperbolic models were designed to account for as well as the facts that are in contradiction with these models, i.e., Keren and Roelofsma's disappearance of the immediacy effect when the present is uncertain. In the words of the entanglement thesis, Halevy's uses of formal language articulate the non-standard theoretical conventions of decision under uncertainty to obtain the behavioral implications that account for facts that the standard and non-standard theoretical conventions of decision over time cannot altogether account for. Furthermore, Halevy justifies this use of formal language by, among other references in economics, arguments from both *P*sychology and from the *p*syschology he shares with decision makers. From *P*sychology, he relies on Keren and Roelofsma's experimental results against

Rachlin (Halevy 2008, p.1155 explicitly takes a stance against the latter). From *p*sychology, he follows the argument from the possibility of sudden death which Keren and Roelofsma call an "analytical argument" (1995, abstract and pp.293-4) by which they mean an argument from common sense and not from experimental results.[38]

The contribution on the quantitative issues underlying the interactions across dimensions is Steffen Andersen, Glenn Harrison, Morten Lau and Elisabet Rutström's (2008) "Eliciting Risk and Time Preferences", published in the *Econometrica* (see also their 2014a; b). These authors argue that the great diversity and variance in measured discount rates mentioned in the previous section is mainly due to the assumption of linear utility in the majority of experimental work eliciting discount rates, e.g., assuming $u(x_t) = x_t$ when computing utilities between now and the next period in $u(x_1) + \frac{u(x_2)}{1+\rho}$. The interaction with risk preferences comes from their interpretation of $u(.)$'s curvature through EU, i.e., as representing risk attitudes (2008; 2014a contains an interpretation through RDU). They accordingly develop an experimental methodology where decision makers' risk attitudes are systematically estimated from risky decision problems and then used in the estimation of time preferences from intertemporal decision problems. Their main result is that, under exponential discounting, the estimated average discount rate is significantly lower and more stable than usual (2008; 2014a). And under hyperbolic discounting, the non-constant decrease of the rate of discounting is smaller when concave utility is estimated through risk preferences than it is when assumed to be linear (2008). From the general perspective of this dissertation, notice that Andersen et al.'s contributions nicely illustrate the entanglement of facts and theoretical convention in the classical sense of 'theory-ladeness', i.e., how the facts about observed discount rates are laden into the theoretical conventions used to make them observable.[39]

From the perspective of this chapter, two aspects of Andersen et al.'s contribution are worth noting. The first one is the methodological debates it has triggered with other experimentalists, arguing notably against its rather explicit primacy of risk preferences (Andreoni and Sprenger

[38] Besides the paper by Epper and Fehr-Duda (2015) mentioned in the previous subsection, Baucells and Heukamp (2012) another *theoretical* contribution of the post mid-2000s' interactions across the dimensions of risk and time. Saito (2011) provides a non-trivial technical correction to Halévy (2008).

[39] They even statistically reject the existence of quasi-hyperbolic discounting from their (2014a) data, but it should be noted that none of their intertemporal experiments involve tradoffs between Now and Later, only between consequences that are Later or in The Future.

2012; 2015; Abdellaoui et al. 2013; Wakker[40]; and esp. Andreoni et al. 2015). These other experimentalists argue that it is not clear why the concavity of $u(.)$ expressing risk attitudes in atemporal settings should alone and adequately account for its concavity in temporal settings. The second aspect is their econometric use of so called 'mixture models', which, roughly, allow to analyze a data set as if the choices were generated not by one optimization process but two (see 2014b for a general discussion). These two processes can be taken from different models, e.g., by EU *and* prospect theory (in Harrison and Rutström 2009) or by EDU *and* hyperbolic discounting (in Andersen et al. 2008, sect.3.D; 2014a). Or these can be taken from one model which postulates two such processes (in Andersen et al. 2008). In this latter case, Andersen et al. justify this methodological strategy through arguments from *p*-&-*P*sychology. They argue that it is well in line with new economic models (Benhabib and Bisin 2005; Fudenberg and Levine 2006) inspired from *P*sychology that can rationalize *classical* challenges from behavioral economics (i.e., from the previous section) within the dimension of risk and time. These models are the 'dual models' that are at the center of next chapter, where we shall see how the ones from Fudenberg and Levine can also rationalize challenges about interactions *across* dimensions. However, though the latter has indeed the unificatory power suggested by Andersen et al., there is a latent primacy of time preferences in their work that is arguably not in line with Andersen et al.'s work.

Summing up, interactions across the three dimensions have triggered a set of contributions where the issue of the primacy of one type of preference over the two other is an *explicit* issue. Though the primacy of risk preferences is still clearly in place (Andersen et al.) and can be justified from *p*-&-*P*sychology (Halevy). The changes in the entanglements of facts, values and theoretical conventions documented earlier in this section tends to undermine this primacy, suggesting a primacy of time preferences instead. Again, we emphasize that the marking of linguistic distinctions within and across Rastier's table is a condition of possibility for the intelligibility of the underlying theoretical and empirical issues.

[40] See Wakker's annotated bibliography available at http://people.few.eur.nl/wakker/refs/webrfrncs.docx, last consultation: 21/01/2016.

Conclusion

This section has provided an empirical and theoretical picture of the main challenges posed to *both* behavioral and standard economics by experimental results on the interactions across the three dimensions. Though the primacy of risk over time over social preferences can still be felt in these contributions, a primacy of time preferences undermines it. Notably regarding the positive/normative issue, value judgments from dynamic consistency are being increasingly made to argue for the rationality of the factual violations of EU's theoretical conventions in the dimension of risk (following Machina's pioneering contribution) and of other people. We have seen one argument from *P*sychology against a primacy of time preferences, namely that THE FUTURE is necessarily UNCERTAIN because of the omniPRESENT possibility of future death while the PRESENT can be certain when consequences are occurring IMMEDIATELY. We have however suggested that a grounding of a primacy of time preferences from *P*sychology may be available in the dual models used by Andersen et al. Anticipating on the next chapter, these models provide another path to theoretical unification that the one offered in the theoretical contributions discussed in this section, notably because the issue of interdisciplinary between economics and *P*sychology is more prominent in the former. Finally, as in the previous section, one condition of possibility for the discussions provided here is the marking of linguistic distinctions within and among Rastier's table.

Conclusion and transition: from three dimensions to two sub-individual entities

This chapter proposed to see rationality in 3D. That is, whether in the classical challenges within dimensions or in the recent ones across dimensions, a systematic discussion of the three dimensions *altogether* has been conducted for the sake of a better understanding of the behavioral *versus* standard economics debates. Both the positive/normative issue and the issue of interdisciplinarity underlying behavioral economics, it was argued, are better understood when seeing rationality in 3D. The use of Rastier's typographical convention of capitalizing instances of his table presented in the introduction of this chapter was meant, not only to see rationality in 3D more vividly, but also to see that one condition of possibility for these challenges

to be even intelligible is the linguistic marking of the distinction within and across his table. Indeed, if a decision modeler cannot do that, then, e.g., an experimenter cannot implement the experiments discussed in this chapter. And if the decision maker cannot recognize that, then all the results discussed in this chapter loose some of their meaning, at least in their normative implications. We also saw a progressive reversal of the primacy of risk over time over social preferences towards a primacy of time over risk over social preferences.

It can be argued that seeing rationality in 3D has two main and related implications. Firstly, an explicit motivation shared by most contributions on interactions across dimensions – which, recall, focus on distinct *pairs* of dimensions – is that, in the real world, the two dimensions they treat are always in interactions. Seeing this world though a one-dimensional theoretical lens, it is usually argued, leads to significant biases in empirical observations. Hence taking all these contributions together, i.e., seeing rationality in 3D, implies that interactions across *the three dimensions* are always at play. This is at least plausible for important economic decisions, so that the three dimensionsal picture presented here may be used to critically reinterpret existing empirical studies that used an unidimensional theoretical framework for interpretation. This first implication will not be investigated in this dissertation (together with Cléo Chassonnery Zaïgouche and Judith Favereau, we try to do so in a working paper). Secondly, it is by seeing rationality in 3D that one can recognize that the whole empirical picture presented in this chapter points a non-trivial area of reconciliation between behavioral and standard economics. That is, *accounting for interactions within and across the three dimensions* suggests itself as a relevant criterion for the methodological evaluation of claims of theoretical unification, and by the same token of reconciliation, between behavioral and standard economics. In the next chapter, we will use this criterion to scrutinize a set of models – the 'dual models' mentioned at the end of this chapter – that claim to achieve such unification and (maybe to a lesser extent) reconciliation, mainly by representing the individual economic agent as the interaction of two sub-individual entities.

Chapter 3

The rise of dual models: behavior as language

"With explanatory power comes explanatory responsibility"
(Rabin 2013, p.622)

In the previous chapter, we saw that an area of convergence between behavioral and standard economics has recently emerged, and lies in the interactions across risk, time and social preferences. At about the same period, i.e., from the mid-2000s onwards, prominent behavioral economists started, in their methodological manifestos, to make strong claims about a new way of modeling individual behavior. The latter consists in representing behaviors of individual economic agents as the outcome of an interaction between two sub-individual entities. The strong claims concerned how these models could play an important role in the convergence of behavioral economics and standard economics, both between themselves and with other cognitive and behavioral sciences. The clearest statement of this is by Camerer, who is worth quoting at length because I will then use his words to set the goals of this chapter:[1]

> "Tremendous progress has been made in going from deviations and anomalies to general theories, which are mathematical and can be applied to make fresh predictions. [...] Excluded from ["rational-choice principles and behavioral economics

[1] See also Camerer, Loewenstein and Prelec (2005, p.43, p.56), Ashraf, Camerer, and Loewenstein (2005, p.132), Camerer (2006, pp.199-200; 2008, p.55, pp.62-63), Loewenstein, Rick, and Cohen (2008, p.650 and p.665), Angner and Loewenstein (2012, p.676).

> alternatives"...], and from the basic ideas ["and tools of behavioral economics"...] are a rapidly emerging variety of formal "dual system" models, drawing on old dichotomies in psychology. These models generally retain optimization by one of the systems and make behavior of another system automatic (or myopic) and nonstrategic so that extensions of standard tools can be used. (Intuitively, think of part of the brain as optimizing against a new type of constraint – an internal constraint from another brain system, rather than a budget constraint or an external constraint from competition.) In Kahneman (2003) the systems are intuitive and deliberative systems ("systems 1 and 2"). In Loewenstein and O'Donoghue (2004) the systems are deliberative and affective; in Benhabib and Bisin (2005) the systems are controlled and automatic; in Fudenberg and Levine [2006] the systems are "long-run" (and controlling) and "short-run"; in Bernheim and Rangel [2004] the systems are "hot" (automatic) and "cold". In Brocas and Carillo [2008] a myopic "agent" system has private information about utility, so a farsighted "principal" (who care about the utility of all agents) creates mechanisms for the myopic agents to reveal their information. [...] *In the years to come*, careful thought will probably sharpen our understanding of the similarities and differences among models. [....] And of course, empirical work is needed to see which predictions of different models hold up best, presumably inspiring refinements that might eventually lead to a single model that could occupy a central place in microeconomics." (2006, pp.199-200, my emphasis)

The years have come. Does "careful thought" have since "sharpen our understanding of the similarities and difference among" the "formal "dual systems" models"? Is there "a single model that" now occupies "a central place in microeconomics"? In a less dramatic tone: nearly a decade has passed, and that seems enough to make these two questions legitimate. The main goal of this chapter is to answer these two questions by scrutinizing the limits and scopes of these models, how they are to be situated between economics and *P*sychology and how the positive/normative issue is articulated within them.

There have been some thoughts on the similarities and differences among these models (see e.g., Ross 2010; Ross, Ainslie and Hofmeyr 2010; Ainslie 2012; Alós-Ferrer and Strack 2014). But these have been made from fairly narrow angles of analyses, and none takes the five models *altogether*. This chapter therefore proposes a more complete comparative exercise in order to further our understanding of these models. It ends up suggesting that there is not yet a single (new) model for microeconomics, but an increasing number of (standard) economists are seriously thinking about it. We can derive a simple criterion to frame and evaluate the answers to these two questions from the previous chapter: can these models account for interactions within and across risk, time and social preferences? One of the main conclusions of this chapter is that Fudenberg and Levine's model is the only one that accounts for interactions *across* dimensions.

This makes it the most promising candidate for the position of new central microeconomic model, at least from the perspective of this dissertation.

When the 'strong claims' are taken altogether (i.e., Camerer's quote and the references in the previous footnote), there is a fairly clear agreement that five models have independently implemented this new form of modeling in the mid-2000s: Loewenstein and O'Donoghue's (2004), Jess Benhabib and Alberto Bisin's (2005), Bernheim and Rangel's (2004), Fudenberg and Levine's (2006), and Isabelle Brocas and Juan Carillo's (2008). In these papers, as in the literature that references them, we do find echoes to Camerer's observations: a new way of modeling individual behavior seems to have emerged in economics which may achieve theoretical unification not only between behavioral and standard economics, but also between economics and other behavioral and cognitive sciences. There is no label to refer to these models. 'Dual-*processes* model', 'dual-*systems* model' and 'dual-*selves* model' are used somewhat interchangeably across and within the strong claims, across and within the models themselves and across and within the literature that references them. As we will see, this terminological diversity is important in itself. It provides some clues as to how these models broadly fit in the relations between economics and *P*sychology. But for the sake of convenience, I will nevertheless refer to these five models by using a common label: *dual models* – and their authors, the *dual modelers*. Beyond convenience and sobriety, 'dual' and 'models' capture two sets of features of these models. Some brief comments on these two words are therefore in order, which I will do by presenting the main conclusions that will be drawn in this chapter.

What kind of 'models' are dual models? Both 'dual processes' and 'dual systems' refer to what can be called a 'dual trend' in various areas of *P*sychology. It can be argued that this trend is in fact well known outside of *P*sychology because a highly representative contribution to it is Kahneman's (2003; 2011) famous System 1/System 2 framework (roughly, the systems are two sets of processes responsible for our irrational and rational tendencies, respectively). This trend is driven by *theoretical unification*, in the sense of a search for one framework to fit several phenomena (i.e., experimental regularities), theories and models. Dual models are likewise explicitly driven by theoretical unification, but differently. On the one hand, they are concerned with phenomena, theories and models that are at the center of the 'behavioral economics *versus* standard economics' debates. On the other hand, uses of formal languages are

central in such theoretical unification, whereas it is not so in the dual trend in *P*sychology. This theoretical unification is achieved through various *applications* of dual models, meant mostly as contributions to economics. By contrast, in constructing the *cores* of their dual models, most dual modelers, who, recall, are *economists*, intend to make a direct contribution to *P*sychology. The issue is of interdisciplinarity is however tricky here because these contributions are done without dual models being influenced by the dual trend in *P*sychology (e.g., by an existing dual framework from *P*sychology). Unlike dual models, Kahneman (2003; 2011) interprets much of his work through the lenses of an existing dual system in *P*sychology (Stanovich and West 2000), and does not use any formal language (it proposes a table articulating some crucial conceptual distinctions). Hence his work will not be discussed on a part with the other dual models in this chapter. Finally, 'Dual *selves*' is a label related to, though different from, 'multiple selves' models economics (e.g., Laibson 1997 discussed in the previous part). The difference is that all selves of a multiple selves model are equal in the sense that they are all the exact same type of entities, i.e., just 'selves', while in dual selves models there are two types of selves, one focusing only on the present and one focusing on both the present and the future. Furthermore, dual selves models explicitly claim to be alternatives to multiple selves models.

Though they have slightly different backgrounds from economics and *P*sychology, dual models share a central conceptual feature, related to the qualifier 'dual'. So in what sense are dual models 'dual'? Dual models consist in modeling *individual behavior* as the outcome of the *interactions* between *two* sub-individual entities, which can be two processes, two systems, two selves, etc. – 'entities' is here used to remain as general and neutral as possible. These two entities are different from, though not unrelated to, (1) the two main sub-individual *entities* in the standard approach, namely preferences and beliefs, and (2) the *multiple* sub-individual entities in the multiple selves approach. In a dual model, the interactions between its two sub-individual entities represent human's capacities to exercise what is usually called in both everyday and scientific ordinary language *self-control* or *willpower*. That is, our capacities 'to restrain our behavioral impulses', 'to resist temptation', etc. The outcome of the interactions between the two sub-individual entities, dual modelers argue, *is* the behavior of an individual entity, namely a human economic agent. In a way, dual models update old philosophical questions about the duality Man and instead of representing the agent as more driven by one or the other side of

that duality, they try to show how the agent is created by their interaction. We shall however not be concerned with these philosophical issues in this chapter, to focus instead on issues raised by dual models in terms of economic analysis.

The positive/normative issue in dual models is not discussed explicitly by dual modelers (with the exception of Bernheim and Rangel). The issue is ambiguous because, there are two ways of interpreting the normative dimension of dual models. On the one hand, one can equate rationality with one sub-individual entity and any deviation from that cannot be normatively justified. On the other hand, one can equate rationality with the decision maker's evaluation, but not to any of its underlying sub-individual entities (one is 'too' rational, while the other is irrational). This latter interpretation is in line with a predicament often made by philosophers talking about normativity, namely that 'ought implies can'. In economists' language, self-control is inevitable, they are a constraint (albeit an internal one) under which the decision maker ought to optimize. This is Camerer's implicit interpretation in the opening quote of this chapter. But we shall see that the constraint *is not* one of the sub-individual entity, as Camerer suggests, it is rather the medium by which two different optimization processes (namely the two sub-individual entities) interact. [2]

We have just stated what this chapter aims to clarify in terms of both the issue of interdisciplinarity and the positive/normative issue within dual models. The role of language in economic rationality will here be tackled by abstracting from the uses of language by economic agents. Hence we focus solely on the uses language from dual modelers. More precisely, the goal here is to scrutinize a set of technical language borrowed from *P*sychology and mixed with the formal language of economics to represent individual behaviors. In a sense, individual behaviors are taken by dual modelers as a language to be decoded and the decoding is operated through the mix of technical language from Psychology and formal language from economics. Hence the second part of the title of this chapter, *behavior as language*.

This chapter is structured in three sections. *The cores* of dual models are presented through an intuitive example and by marking the uses of technical language from *P*sychology, which is

[2]The two ways of interpreting the normative dimension of dual models could be grounded further into the literature on normativity and rationality in the cognitive and social sciences (see resp. Stein 1996; and Turner 2010). However, the vocabulary and connected issues of these literatures (esp. in Turner 2010) would render the comparative examination of dual models way too complicated. This chapter can however be seen as a first step for such further work.

then put in perspective with how *the backgrounds of dual modelers as economists* affect their respective uses of formal languages (3.1). This first subsection is mainly concerned with the interdisciplinarity issue in an intuitive fashion, i.e., the relation between economics and *p*sychology within dual models. The positive/normative issue within dual models start to be discussed from the second section onwards. *The shared backgrounds of dual models* from *P*sychology (i.e., the origins of the technical language marked in the first subsection) and from economics (i.e., the literature in which they seek to contribute) are presented to further our understanding of the similarities and contrasts of their intended contributions (3.2). The applications of dual models are compared, which is tantamount to comparing how they achieve *theoretical unification* within and/or across the three dimensions discussed in the previous part (3.3). The conclusion of this chapter summarize the arguments and emphasize on one common limit to dual models, namely their inadequacy to capture so-called '*framing*' phenomena.

3.1 Cores: *P*sychology and the backgrounds of dual modelers

The goal of this subsection is to compare the cores of dual models. The issue of interdisciplinarity is discussed intuitively by marking the technical language from *P*sychology used by dual modelers and explaining it through a common decision example constructed for this purpose (3.1.1). That is, the details about the scientific meaning of this language as well as its origins from various areas of *P*sychology is silenced (because it will be scrutinized in the next section). Dual modelers' uses of formal languages is then discussed informally by emphasizing their respective backgrounds as economists (3.1.2).

3.1.1 The cores of dual models with a piece of cake and a cup of coffee

Imagine you are at a conference and there are delicious cakes during the coffee breaks. At a certain break, and without thinking too much about it, you are about to take a piece of cake and then a cup of coffee. But at the last second before you reach the piece of cake, you think you should not get cake and coffee for some reasons, e.g., you are on a diet and have had too much coffee today. This thought makes you slightly deviate the aim of your hand from the cake to grab a cup of coffee, which you drink. So you initially wanted both cake and coffee, and then

wanted neither cake nor coffee, to end up choosing just coffee. In this subsection, we shall see how the cores of the five dual models can or cannot explain this simple example. Notice that the structure of the example is different from the more traditional intertemporal problems discussed in the previous chapter (e.g., small consequence now vs. bigger consequence later). Indeed, the specificity of its structure is meant to reflect the specificity of dual models with respect to the models of intertemporal choice discussed in the previous chapter. Throughout this section, I will use single inverted comas ('like this') to mark the specific terms from *P*sychology used by dual modelers, the exact references of which will be discussed in the next section; I will also italicize (*like that*) the specific terms that are proper and central to the respective dual models.[3]

The two sub-individual entities at the core of Loewenstein and O'Donoghue's dual model are two 'systems', a *deliberative* one and an *affective* one. These two systems are two sets of psychological and possibly but not necessarily neural 'processes', that are supposed to be at play in the various conflict we encounter in everyday decision making. Loewenstein and O'Donoghue postulate that each system have their own optimal consequence. In our example, the *affective optimum* is to take a piece of cake and a cup of coffee, while the *deliberative optimum* is to take nothing. The actual consequence, i.e., taking just coffee, is the outcome of this conflict, which depends on three parameters in Loewenstein and O'Donoghue's dual model. One parameter is 'willpower strength'. It represents the mental effort needed to push behavior towards the deliberative optimum at the expense of the affective system. Intuitively, it represents how disagreeable it was when you steered your hand away from the piece of cake towards the cup of coffee. The more disagreeable it feels, the more likely it is that you take the piece of cake. Another parameter is 'cognitive load'. It represents the amount of thinking you dedicated to something else than the conflict. The higher the cognitive load is, the more likely it is that you take the piece of cake. A very high cognitive load can be illustrated by a telephone number you are trying to memorize during the conflict: you barely notice this conflict and find yourself eating the cake before you know it. A lower cognitive load can be illustrated by a mildly interesting conversation you are having during the conflict: you can still think about your diet and steer your hand towards the cup of coffee. The last parameter is the intensity of affective 'motivations'.

[3] I thank Alberto Bisin for having pointed out to me the difference in structures between the problem presented here and the ones from the previous chapter.

It represents how strongly your affective system is attracted by a given object of choice in a given situation. The stronger the intensity of affective motivation is, the more likely it is that you take the piece of cake. To illustrate, such intensity is very high for a cake you particularly like, which is very well presented, and that you can have here (not in another room) and now (not later as if you had to chose beforehand what you will be eating during coffee breaks). By contrast it is lower for a cake you don't particularly like, or for one that you particularly like but which is not well presented or is not available here and now. Giving values to these three parameters allows a utility calculus of the cost of willpower that the deliberative system pays to lower the affective system's utility (by not taking cake and coffee), which also lowers (since it is a utility cost) the former's utility. The core of Loewenstein and O'Donoghue's dual model has no problem accounting for the kind of in-between compromise of our example. The cost in utility to be paid by the deliberative system to take nothing was higher than the cost in utility to be paid to not take cake plus the loss in utility for taking coffee.[4]

The two sub-individual entities at the core of Fudenberg and Levine's dual model are two types of selves, a *long-run* one and *short-run* ones, playing a sequential game. The same long-run self plays during all the sequences of the game, while every short-run selves plays only one sequence. One sequence of the game consists in two stages: the long-run self first plays a *self-control action* which influences the current short-run self's utility, who then plays an *action*. All short-run selves optimize regardless of history and future, and the long-run self knows that and plays his self-control actions accordingly. In our example, your long-run self played a self-control action such that the current short-run self's utility for cake and coffee becomes inferior to the utility of just coffee, whereas the reverse would have been the case without any self-control actions. So why didn't your long-run self play a self-control action such that taking nothing becomes the optimal action for your short-run self? In Fudenberg and Levine's dual model, the long-run self's utility is the exponentially discounted and expected sum of all the short-run selves' utilities. Hence losses of utility from self-control actions are incurred by the long-run self as well. Thus, in our example, your long-run self decided that the cost in utility to be imposed so that your short-run self takes nothing was too big, and compromised for taking just coffee.

[4]It could be argued that such psychological interpretations of utility is at odds with standard economics. As already pointed in the previous chapter (§2.1.2), this is not at odds with the interpretation of utility in behavioral economics; this remark also applies to dual modelers.

There is one non-trivial refinement of the core of their dual model, with short*er*-run selves playing more than one sequence (Fudenberg and Levine 2012b; c). In this case, the long-run self commits his self-control actions to all the sequences played by one shorter-run self, i.e., he cannot change his self-control actions between these sequences.

The two sub-individual entities at the core of Benhabib and Bisin's dual model are two types of processes, 'controlled processes' and 'automatic processes'. The core of their dual model is formalized in a two period intertemporal decision problem, where the first period is not necessarily now. The more the first period is in the future, the more controlled processes are able to evaluate the utility of the consequences through exponential discounting. And the closer to the immediate present is the first period, the more likely it becomes that automatic processes evaluate the utility of the present consequences as disproportionately greater than the utility of future consequences. In our example, the first period is immediate: cake and coffee are to be consumed immediately. The second period can be thought of as everything that happens after, e.g., the next coffee break and/or tonight's cocktail party. Controlled processes evaluate the utility of taking nothing as superior to the utility of cake and/or coffee; automatic processes evaluate the utility of cake and coffee as superior to just coffee or nothing. *If* we postulate these two optima for these two processes, as we did with the other dual models, *then* Benhabib and Bisin's dual model cannot account straightforwardly for the decision to just take coffee. This is so because they formalize one function of controlled processes, namely the total inhibition of automatic ones. Such inhibition is called 'cognitive control', and is said to occur through the 'supervisory attention system'. Benhabib and Bisin model self-control as an instance of cognitive control through an *attention cost parameter* and this cost is measured in utility. The attention cost parameter represents the utility cost above which exerting cognitive control will yield less utility to the decision maker than the utility given by automatic processes' optimum. In our example, focusing attention to the consequences in next break and/or tonight's cocktail party diverts attention from taking cake and coffee now (in order to take nothing now), which represents a cost in utility. If this cost is superior to the difference between (1) the utility of taking nothing now and something later and (2) the utility of taking cake and coffee now but nothing later, then cognitive control is not exerted, i.e., automatic processes are not inhibited hence cake and coffee are consumed now. If the cost is inferior to that difference, then cognitive

control is exerted, i.e., nothing is consumed now. It should be emphasized that Benhabib and Bisin could easily explained our example *if* we had postulated different optima for the processes, e.g., with just coffee as the controlled processes' optimum. However, doing so makes the comparison with other dual models less direct (the same remark applies to Bernheim and Rangel's presentation below).

The two sub-individual entities at the core of Bernheim and Rangel's dual model are two systems, 'the mesolimbic dopamine system' and a *cognitive forecasting system*. The mesolimbic dopamine system is a neural system, sometimes called the reward system. It generates instantaneous utility, notably based on the association of 'cues' in one's present environment with past short-term hedonic experiences: a mechanism that Bernheim and Rangel call the *hedonic forecasting mechanism*. In our example, the cues can be the smell of coffee, or empty cups of coffee, or empty plates of cake; and the hedonic forecasting mechanism is responsible for wanting to take a piece of cake and a cup of coffee. This evaluation of consequences competes with long-term evaluation from the cognitive forecasting system. In our example, this system forecasts the long-term utilities of taking nothing, which it evaluates as higher than taking cake and/or coffee. Bernheim and Rangel model this competition within a stochastic and dynamic intertemporal model of 'addiction'. In the core of their dual model, the decision maker can be in two modes of decision making. In a *hot mode* he automatically chooses an addictive consumption, e.g., cake and coffee; in a *cold mode* he chooses whether or not he does so. At each period, the decision maker faces a sequence of two decision problems. He first chooses a *lifestyle activity*, always in the cold mode. Lifestyle activities have different levels of utility, in inverse proportion to the probability of encountering addictive consumptions, or cues of addictive consumptions. In our example, three lifestyle activities can be: going to the coffee break room (high utility), or going outside but close to the break room (medium utility), or going directly to the room where there will be the next session you want to attend but which is far away from the break room (low utility). Once the lifestyle activity chosen, the decision maker faces a probability of entering the hot mode. This probability depends ultimately on factors, especially past consumptions and cues in the environment, that influence the utility of the addictive consumption as evaluated by the hedonic forecasting mechanism. There is a utility threshold above which that utility completely override the cognitive forecasting system's evaluation, so that the decision maker

enters in the hot mode. In our example, the decision maker may choose to go to the coffee break room, derived substantial utility from cakes and coffee at previous breaks and sees everybody else eating cake and drinking coffee. The probability that he enters in the hot mode is therefore quite high. If it occurs, he consumes cake and coffee (if it would have occurred while in the next session's room he couldn't have consumed cake and coffee, i.e., he precommitted against it); if it does not occur, he chooses whether he consumes cake and coffee or not. If we postulate that the mesolimbic dopamine system's optimum is cake and coffee and the cognitive forecasting system's optimum is taking nothing, then Bernheim and Rangel's dual model is unable to straightforwardly account for the decision to just take coffee.

The core of Brocas and Carillo's dual model is trickier to explain because, unlike all other dual modelers, they do not build a core for their model and then apply it to different problems. It can nevertheless be reconstructed by focusing on what is shared by the great majority of their applications. The two sub-individual entities at the core of their dual model are pairs of 'brain systems' constituted of a *principal* system and an *agent* system. The decision maker can choose how much to do of *two actions*, and the principal system wants to do more of one action and less of the other than the agent system. In our example, the two actions are (1) consuming cake and coffee and (2) consuming nothing. The agent system's optimum is to do only the former, and the principal system's optimum is to do only the latter. The two systems have therefore conflicting preferences over the relative quantity of the two actions. Conflicting but not opposite: the principal derives its utility from the agent's utility, but with different marginal utility. Furthermore, the principal can control the agent's actions, but to do so optimally, she needs information about the agent's marginal utility. Brocas and Carillo capture this in an information parameter, over which there is an asymmetry of information between the two systems: the principal does not know the agent's marginal utility. In our example, the principal does not know how much the agent will actually enjoy eating cake and coffee and dislike eating nothing. The principal can design an incentive mechanism to reveal this information from the agent. With this information the principal can then constrain (as if a contract was designed) the pairs of actions available to the agent, so as to achieve a second best solution from the principal's point of view. Such constraint depends on other characteristics of the principal-agent relation, which are application-dependent. In our example, the two main constraints discussed by Brocas

and Carillo apply quite well: the principal wants to constrain the agent's action for the sake of future marginal utility (to enjoy the next break or tonight's cocktail party better), or because she is aware that the agent's present marginal utility is biased, e.g., once the cake eaten, the agent will not feel as satisfied as he thought he would have been. In both cases, the second-best solution is obvious: just take coffee.

Regarding the methodological perspective of this chapter, a first comparative summary of the cores of dual models can be put as follows. Fudenberg and Levine are the only ones not using any language from *P*sychology in the construction of the core of their dual models. That is, they are only using only *p*sychological language to motivate and/or interpret their theoretical constructs, most of which are fairly standard game theoretical notions. All other dual modelers make explicit references to contributions in *P*sychology, of which we have only marked the technical language through inverted commas, in the construction of the cores of their dual models. The cores of both Loewenstein and O'Donoghue's and Brocas and Carillo's dual models are the only ones that are not inherently temporal, in the sense that they can account for static or atemporal decision problems, while the others would do so by formally bringing a temporal dimension. The core of Bernheim and Rangel's dual model is the only one that is inherently under uncertainty, in the sense that there has to be a specific risk associated with choices of lifestyle activities for the utility of these choices to be meaningful. No dual models' core is inherently social, at least in a sense related to the standard notion of social preferences as discussed in the previous chapter (Bernheim and Rangel's lifestyle activities is however social in the sense of choosing among which kind of people one wishes to interact). Finally, the cores of Bernheim and Rangel's and of Benhabib and Bisin's dual model cannot account straightforwardly for the outcome of our simple example, i.e., just taking coffee. Both Loewenstein and O'Donoghue and Fudenberg and Levine account for it through modeling costs of self-control, while Brocas and Carillo account for it by modeling an asymmetric informational constraint between their two sub-individual entities.

3.1.2 Uses of formal languages from the backgrounds of dual modelers

All dual modelers use formal languages (i.e., mathematics, logic and probability theory) with some slight yet non-trivial differences. These differences are best understood by paying some

attention to the variety of dual modelers' backgrounds as economists. The goal here is to discuss these differences in order to better understand the relation between economics and *P*sychology in dual models. Very broadly these backgrounds can be characterized as: game theory for Fudenberg and Levine, contract theory for Brocas and Carillo, behavioral economics for Loewenstein and O'Donoghue, macroeconomics for Benhabib and Bisin and public economics for Bernheim and Rangel. They will be discussed in this order.

The most central formal feature of the core of Fudenberg and Levine's dual model is the reduction of the game between a long-run self and short run selves to a simple optimization program for a single decision maker (2006, Theorem 1). In economics, the language of 'multiple selves' historically comes from the use of game theory to bypass the technical impossibility of modeling inconsistent intertemporal preferences through straightforward optimization under constraint (cf. previous chapter). The core of Fudenberg and Levine's dual model can therefore be seen as a historically non-trivial contribution to this literature. It provides substantial gains in analytical tractability for economists that want to generate dynamically inconsistent intertemporal preferences. Their expertise in game theory is certainly not innocent here, as can be grasped by their uses of the adjectives 'long-run' and 'short-run' to qualify the two types of self. Both Fudenberg and Levine have made contributions on games between long-run and short run players that were unrelated to multiple selves problems (see the collected papers in Fudenberg and Levine 2009). Considering such a game within an individual allows for assumptions that would not make sense otherwise. Especially, one player cannot control the actions of another player in the same fashion as one self can control another self. This is the key insight that Fudenberg and Levine use in their characterization of a specific equilibrium for their game, which they prove to be equivalent to a simple optimization problem for a single decision maker. The contribution of the core of their dual model is thus strictly internal to economics.

The core of Brocas and Carillo's dual model that can be reconstructed from their applications is also strictly intended as contributions to economics (see 2008a; 2014). Throughout their work, they explicitly use the formal language from "the tradition of the contract theory literature" (p.1317). They more precisely use a blend of contract theory with mechanism design theory (see Börgers 2015, pp.2-3 on how common and natural is this blend). This is well in line with their formations at the *Toulouse School of Economics*, an institution famously known for its researches

on contract theory and mechanism design. Though the substance of the core of their dual model is intended as contributions to economics, they nevertheless recurrently argue that the mathematical techniques of constrained optimization with informational asymmetries are more rigorous and precise than computational models used in neurosciences. Hence, the methodology underlying the core of their dual models is also intended as contributions to neurosciences, including *P*sychology (see esp. Brocas 2012).

Loewenstein and O'Donoghue insist on their main contribution being twofold: (1) the theoretical construction of their affective system and its interactions with the deliberative system, and (2) the formalization of this construction through rather simple functional relations. The core of their dual models is presented as a direct contribution to the dual trend in *P*sychology. Their uses of a simple formal language can be seen as a way to communicate with psychologists, in an effort of integration between economics and *P*sychology. Hence the core of Loewenstein and O'Donoghue's dual model is a formalized instance of the claims sometimes made by behavioral economists about the possibility and potential fruitfulness of an integration between economics and Psychology (e.g., Camerer 1999; Loewenstein, Rick and Cohen 2008), which standard economists tend to criticize (see, e.g., Gul and Pesendorfer 2008).

In a similar fashion, Benhabib and Bisin explicitly present the core of their dual model as a theoretical contribution to the literature in *P*sychology from which they borrow the constructs of automatic and controlled processes. They acknowledge that the formalization they propose has empirical implications at the neural level which have never been tested. They thus propose an experimental design in order to make such a test (see 2005, Fig. 1 p.467). Furthermore, in their contributions to *The Foundations of Positive and Normative Economics* (Benhabib and Bisin 2008), their dual model is taken as an illustration of their methodological reflections on the modeling of individual behavior for positive economics, which prone a rather strong form of interdisciplinarity with *P*sychology. Their backgrounds as macroeconomists do not show up in the core of their dual model, however. It does in their application of the core of their dual model to characterize a representative agent in a dynamic stochastic economy (2005, pp.467-479).

Bernheim and Rangel take great pains to ground their sub-individual entities in the "psychology, neuroscience, and clinical practice" of addiction (2004, p.1582). The core of their dual model is intended as a positive contribution to the modeling of addiction behaviors generally

	FL	LO	BC	BB	BR
Account for in-between choice (e.g., just coffee)?	Yes (utility costs of self-control)	Yes (utility costs of self-control)	Yes (asymmetric informational constraint)	No (because of an all-or-nothing threshold)	No (because of an all-or-nothing threshold)
Inherently temporal?	Yes	No	No	Yes	Yes
Inherently under uncertainty?	No	No	No	No	Yes
Inherently social?	No	No	No	No	No
Contribution in *P*sychology?	No	Yes	Yes	Yes	Yes
Use language from *P*sychology?	No	'systems' & 'processes'; 'willpower'; 'cognitive load'; affective 'motivation'	'brain systems'	'controlled' & 'automatic' 'processes'; 'cognitive control'; 'supervisory attention system'	'mesolimbic dopamine system'; 'cues'

Table 3.1: Comparison of dual models' cores

(pp.1568-1572), i.e., to both the literatures mentioned previously and to the economics of addiction. Their backgrounds as public economists show up in the applications of their dual model (2004, pp.1572-1580), which consist in the analysis of policies related to addictive behaviors. Furthermore, in their contributions to *The Foundations* (Bernheim and Rangel 2008), which reformulates the whole axiomatics of welfare economics to characterize behavioral mistakes, they use their dual model as an "illustration" of the models of individual behavior that can be used in such a framework (see p. 187; this claim is invariably made in their work in normative economics, see, e.g., 2007b, p.470; 2009, p.79; Bernheim 2009, p.36).

This subsection can be summarized in one short sentence. Every dual modelers but Fudenberg and Levine intend their uses of formal language from economics to make a contribution to *P*sychology.

Conclusion

Table 3.1 summarizes the main outcomes of the comparison made in the two subsections of this section (using dual modelers' initial for space reasons).

Two of this chapter's conclusions can already be glanced at. First, by contrast with the

previous chapter, there is a primacy of time over risk and social preferences. This will be strengthened in the third section on dual model's applications. Secondly, the success of Fudenberg and Levine's dual model among economists suggests that the reconciliation between behavioral and standard economics does not imply a strong interdisciplinarity between economics and *P*sychology. This will be weakened in the third section on dual model's applications, where we will see that Fudenberg and Levine make some non-trivial uses of language from *P*sychology. In any case, to the exception of Loewenstein and O'Donoghue, the movement of theoretical unification between standard and behavioral economics through dual models is mainly carried by economists that were considered 'standard' before the 2000s.

3.2 Shared backgrounds: from *P*sychology and from economics

This section scrutinizes the shared backgrounds from Psychology and from economics against which their dual models are best understood. In the first subsection, the types of *P*sychology from which dual modelers borrowed the scientific language marked in the previous section are discussed (3.2.1). In a second subsection, the contributions from economists with respect to which dual modelers place their contributions are discussed (3.2.2). In both cases, the discussions systematically proceed in two steps. A comprehensive account of the contributions is first given only in an intuitive fashion, usually by putting the reader into the shoes of a subject participating to the experiments underlying these contributions. Then an explicitation of the potential sources of normativity underlying these contributions is proposed. It should be noted that neither the terms from *P*sychology nor the contributions in economics discussed in this section are exhaustive of dual modeler's uses of *P*sychology and economics. The contributions selected for the discussion here are however more central and *shared* across dual models than others. Though I shall only discuss explicitly the dual modelers for whom a given reference is central, each of the references discussed below are usually also used by *at least three other dual modelers*. This is why I use the expression '*shared* backgrounds' throughout.

3.2.1 Participating to, but not influenced by, the dual trend and its critics

This subsection scrutinizes the language from *P*sychology used in dual models by giving brief explanations and illustrations of the references underlying the terms marked in the previous section. We shall see that the relation of dual models to what was called the dual trend in *P*sychology (in the introduction of this chapter) is surprisingly tenuous. Dual modelers are inspired by references that are also discussed in the dual trend in *P*sychology, and their dual models are even referenced by psychologists as being part of the trend (e.g., in Keren and Schul 2009). But with the exception of Loewenstein and O'Donoghue, and later Fudenberg and Levine (in a paper co-written with a psychologist), dual modelers neither situate their contributions within that trend, nor do they make direct references to it.

The first references to be discussed dates back from 1977, when cognitive psychologists Walter Schneider and Richard Shiffrin coined 'automatic processes' and 'controlled processes' in a way that has been substantially used in the dual trend (Schneider and Shiffrin 1977; Shiffrin and Schneider 1977). To illustrate what that means, suppose you are in front of a computer. You have to memorize four digits, e.g., {1,5,3,7}, before you see a series of screen with four letters on each screen, except for one screen where a digit replaces a letter, e.g., {A,B,C,D}, then {D,O,P,Z}, then {F,K,1,H}, and then {O,I,Y,T}. For each screen, you have to press one of two buttons: "no" if you don't see any of the memorized digits or "yes" if you see one of them. This is very roughly the kind of experiments conducted by Schneider and Shiffrin, with a great number of variations, e.g., varying the time between each screens or having to memorize letters (unlike in the previous example) which you then have to detect in series of screen with letters (like in the previous example). The main goal of these experiments was to use the data to construct "a general theory of human information processing" (Schneider and Shiffrin 1977, p.3). At the basis of their theory is a qualitative difference between two types of information processing: 'controlled search', e.g., when you search a memorized letter among other letters, and 'automatic detection', e.g., when you detect a digit among letters. In their theoretical account, these two types of processes cannot operate simultaneously, but are triggered under different conditions, especially regarding the type of learning that may have accompanied the memorization of the target set of digits or letters. This distinction between automatic processes and controlled processes has

been influential in cognitive *P*sychology, notably for theories of attention as witnessed in the comprehensive and critical review by cognitive psychologists Stephen Monsell and Jon Driver (2000).

It is from this tradition in cognitive *P*sychology that Benhabib and Bisin borrow the constructs of automatic and controlled processes at the core of their dual model. Notice that in the kind of tasks used by these cognitive psychologists, there are clear correct and incorrect answers. Pressing "yes" when no memorized digit appears on the screen is a mistake in a sense that cannot be said of choices made in behavioral economists' intertemporal choice tasks. In the words of the entanglement thesis, the value judgment of mistaken behaviors is derived from the factual observation of subjects conforming or not to the dual modelers' conventions explicited in the experimental instructions, on the one hand, or to the theoretical convention of dynamic consistency from economics, on the other. Therefore, in the former, the main source of normativity is simply the dual modeler, whose instructions define somewhat arbitrarily what is a correct or an incorrect answer (by contrast with, say, the laws of logic or of probability theory). The distinction between these two types of tasks is related to what has historically been a sharp opposition between 'cognition' and 'motivation' in most of (American) mainstream *P*sychology (see Higgins 2012, p.213); with the notion of 'affect' and 'emotion' being associated with 'motivation'. 'Cognitive control' and 'self-control' denotes the mental efforts one has to do respectively in the former to get right answers and in the latter to make choices that are dynamically consistent. Benhabib and Bisin's contribution to *P*sychology consists in postulating and modeling self-control problems as instances of cognitive control problems. By doing so, they identify dynamic inconsistencies with automatic processes and failures of cognitive control. In a way, their contribution to *P*sychology is to give a formalized cognitivist account of a motivational problem, thus furthering Shiffrin and Schneider's initial ambition of theoretical unification.[5]

Call the type of *P*sychology just discussed 'pure *cognitive P*sychology'. It is in great contrast with so-called '*affective* neuroscience', where the next reference on 'brain systems', and especially

[5] Although they cite Monsell and Driver (2000) approvingly, Benhabib and Bisin seem to be subjected to three interrelated contemporary criticisms reviewed by them. These are the refusals (1) of an automatic/controlled *dichotomy*, (2) of *inhibition* as the only type of interaction between both, and (3) of postulating a unique and unified *system* responsible for all form of cognitive control. We have seen that the core of Benhabib and Bisin's dual model predict that either one of the two processes will be in charge (a dichotomous outcome), through inhibition of the automatic processes monitored by *one* 'supervisory attention system'.

on the 'mesolimbic dopamine system' comes from, namely from the work of Kent Berridge and his co-authors. As Berridge explains in various places, dopamine is a neurotransmitter that, for a long time, was associated with the experience of pleasure (e.g., Berridge 2003; Berridge and O'Doherty 2014). Berridge and his co-authors have argued that the best way to measure the experience of pleasure lies in the observation of facial expressions during the consumption of something sweet (e.g., sugar). It produces "a pattern of tongue protusions, lip sucking, facial relaxation, and the occasional smile" common to humans, monkeys and rodents (Berridge 2003, p.25). Mostly using experiments with rats, they found that high activation of the mesolimbic dopamine system, one of the main brain system involved in the release of dopamine, is *not* associated with the experience of pleasure. Then, to understand with what it was associated, they conducted further experiments where rats were trained to make efforts by pressing a lever to get sugar on some days, and on other days they were conditioned (à la Pavlov) by a red light or a specific sound which occurred just before they receive sugar without making any efforts. Once the conditioning is successful, those sounds or lights are said to be 'cues' for sugar. In the experimental session, amphetamine is injected in the brain of some rats, which triggers a high activation of the mesolimbic dopamine system. The rats can press the lever (they learned that efforts in doing so is rewarded with sugar), the lights or sounds sometimes occur (the cues triggering expectation of sugar without efforts), but in any case no sugar is delivered throughout the experiment. What they observe is that rats that have not been injected amphetamine smoothly decrease their efforts during the session, with no big difference when a cue occurs. Rats that have been injected amphetamine display the same tendency, except when a cue occurs, which triggers a level of effort that is more than four hundred percent higher for about 30 seconds. For Berridge, the main theoretical goal behind these types of experiments is to argue for a neurologically founded distinction between *liking* (the experience of pleasure) and *wanting*, with a sub-distinction between ordinary wanting and cue-triggered 'wanting' (which Berridge always distinguish by uses of inverted commas). Ordinary wanting is akin to "a conscious desire for cognitively-represented outcome" (p.27), or "cognitive wanting" for short; it is the (possibly discounted) anticipation of pleasure. On the other hand, 'wanting' is akin to "cued attraction to salient incentives", or "cue-triggered 'wanting"' for short (ibid). The main contribution from Berridge's work is the identification of the role of dopamine for

'wanting' but not necessarily for wanting and liking.

It is mainly with respect to this tradition from neurosciences that Bernheim and Rangel situate their contributions, especially in two respects. First, like Berridge and many others, the relevance of these findings are used to understand human behavioral phenomena related to addictions. More precisely, Bernheim and Rangel provides a theoretical unification of several such phenomena based mostly on Berridge's work. Second, again like Berridge, they explicitly see choices based on 'wanting' as clearly irrational, by contrast with choices based on wanting. It should be noted that the Berridge's own normative position follows explicitly what we have called earlier the primacy of time preference in the work of Kahneman (cf. chapter 1). In the words of the entanglement thesis, the value judgments of mistakes and irrationality are derived from the theoretical conventions of hedonism together with the factual observation of animal and pathological behaviors.

Next, I would like to discuss the references in *P*sychology related to 'cognitive load' and 'willpower' altogether, and starting with only an illustration of the empirical phenomena they refer to. Dual modelers have associated 'cognitive load' with the experimental results of Baba Shiv and Alexander Fedorikhin (1999). To put yourself into their subjects' shoes, imagine that you are asked to remember either two digits (low cognitive load) or seven digits (high cognitive load), and then to make a choice between chocolate cake and fruit salad. According to Shiv and Fedorikhin's results, there is a strong probability that you will choose cake over fruits with high cognitive load and the reverse under low cognitive load. 'Willpower', on the other hand, is associated by dual modelers to the work of Roy Baumeister, a typical experiment of whom goes as follows (2003, sect. 3.3). Imagine that you enter a room with a table on which there are two plates, one with radishes and one with freshly baked cookies. In a first phase, your task is either to eat only radishes (so that you have to use your willpower not to eat cookies) or to eat whatever you want (so that you don't have to use your willpower). In a second phase, regardless of what you were asked in the previous task, your task is now to solve a problem that, unknown to you, is unsolvable. According to Baumeister's results, there is a strong probability that if you had to use your willpower in the first phase (you were asked to eat only radishes), then you will spend less time (using less willpower) trying to solve the problem in the second phase.

Theoretically, Shiv and Fedorikhin's (1999) goal is to provide evidence for models in *P*sychology

that postulate an impact of cognition on affects (or 'emotion' or even more broadly on 'motivation'). And Baumeister's experiments are meant to provide evidence for his own model of self-control that postulate such an impact. In part, the core of Loewenstein and O'Donoghue's dual model uses Shiv and Fedorikhin's results to develop Baumeister's model. To understand the latter, another empirical tendency from his work needs to be stated: "people who performed minor exercises in self-control for two weeks, such as trying to improve their posture (i.e., stand and sit straight), showed improvements in self-control on laboratory tasks [e.g., spent more time on the unsolvable problem] relative to people who did not exercise" (Baumeister 2003, p.12). Hence Baumeister's 'strength' model conceptualizes 'willpower' or 'self-control' (terms he uses interchangeably) as a muscle: "just as a muscle becomes tired when exercised but eventually becomes stronger, the capacity for self-control becomes depleted in the short run but may be strengthened in the long run if it is used regularly" (ibid). The depletion of willpower capacity observed in Baumeister's work is becoming popular throughout *P*sychology under the name *ego depletion* (see, e.g., Kahneman 2011). The 2012 extension of the core of Fudenberg and Levine's model does formalize ego depletion, unlike both their 2006 initial dual model and Loewenstein and O'Donoghue's (though they roughly mention how it could). Indeed, Loewenstein and O'Donoghue simply add a cognitive load parameter which mediates the use of willpower capacity: for a given task, the higher the cognitive load, the more willpower will be needed. In terms of the normative dimension of these behavioral phenomena, Shiv and Fedorikhin (1999, p.289) cite some of their preliminary results to contrast "[t]he traditional view of impulse behavior as being irrational" from researcher's perspective with the "consumers"' own perspectives, who "do not seem to view impulse behavior as normatively inappropriate, at least immediately after the behavior occurs". They however suggest that this can be mere post-hoc rationalization, not to be trusted by researchers. Baumeister holds a more radical view: failures of self-control due to ego depletion are simply and straightforwardly irrational, and this is justified mainly by appealing to the reader's intuition through casual examples (see esp. Baumeister 2003; 2005; 2015). In both cases, the value judgment of irrationality comes from the factual observation of the impact of cognition on motivation together with the social conventions used by psychologists in their introspection and casual empiricism.[6]

[6]It should be noted that the phenomenon of ego depletion is currently under a replication crisis in *P*sychology,

In one way or the other, the references discussed so far are mentioned in the dual trend in *P*sychology but they are far from being central in it, with the exception of Schneider and Shiffrin's work. The dual trend in *P*sychology arose from three subareas that are not part of dual modeler's shared background from *P*sychology: the 'cognitive *P*sychology of *deductive* reasoning', the 'cognitive *P*sychology of *inductive* judgment', and 'social cognition'. They respectively study the cognitive processes underlying the making of deductive inferences, inductive inferences, and social perceptions and behaviors. Three brief remarks are worth making in order to contrast dual model's shared background from *P*sychology (references are in the next footnote). First, the dual trend is a movement of theoretical unification within which the uses of 'systems' as a set of 'processes' has stabilized around the famous labels *System 1* (including 'conscious', 'fast', 'automatic', etc. processes) and *System 2* (including 'unconscious', 'slow', 'controlled', etc. processes). Second, critics to that trend have targeted two ambiguities in ordinary language uses. One ambiguity concerns the substance of the processes (e.g., 'automatic' and 'controlled'), with many authors meaning different things by the same terms. The other ambiguity concerns the conceptual relation between processes and systems: processes from a given system do not have necessary connections, two processes from each systems can be active at the same time, and interactions between two processes of a given pairs (e.g., 'automatic' and 'controlled') suggest that the binary entities lie on a continuum rather than forming a qualitatively dichotomous distinction. Defenders of the trend have argued that dual system models are still useful at the metatheoretical level, and that the ambiguities are not so pervasive in a given applied work. In other words, they argued that the pairs of binary distinction are neither dualism nor dualities, but ordinary distinction useful from applications to applications. To anticipate on this chapter, we shall argue that this is also the best interpretation that dual models can receive. Third, the sources of normativity in these three subareas, namely *logic*, *probability theory* and *social norms*, are different from the ones spotted above in this subsection.[7]

---

i.e., a set of psychologists have claimed and shown that some underlying effects that were thought to be robust during the past decade are actually not so. Discussing this along with its non-trivial implications for the rise of dual models in economics is beyond the scope of this chapter. Michael Inzlicht is one of the leading psychologist from this group and provide recent update on his scientific blog http://michaelinzlicht.com/getting-better/, last consulted 18/04/2016. A comprehensive platform of replications in *P*sychology is http://curatescience.org/#section-3, last consulted 30/04/2016. I thank Tania Ocana for having pointed this to me.

[7] On the dual trend as a movement of theoretical unification across three subareas of *P*sychology, see esp. Stanovich and West (2000), Evans (2004; 2008), Frankish and Evans (2009), Frankish (2010), Bonnefon (2013);

Finally, one last reference from *P*sychology that we have already encountered in the previous chapter should be mentioned: Ainslie. With the exception of Bernheim and Rangel, Ainslie is referenced by all dual modelers but it would be an exaggeration to count him in dual model's shared background. This is because they reference Ainslie only when they claim that intertemporal preference reversals are pervasive, but they do not borrow much (if at all) theoretical insights from his work. Similarly to the dual trend, it is worth briefly commenting on Ainslie's work for the sake of contrasting dual models' shared background from *P*sychology. Ainslie's (1992; 2001; 2012) work is encyclopedic on self-control in two senses. Empirically, he reviews an impressive amount of dynamically inconsistent human and animal behaviors. And theoretically, he situates his theoretical contribution, which aims at providing a unifying account of these behaviors, within a wide scope in *P*sychology, philosophy and economics. Roughly, the specificity of Ainslie with respect to the other psychologists discussed above is his (partial) inspiration from 'behaviorist *P*sychology'. This is manifest in two features. First, his theoretical account emphasizes the role of 'motivation' and downplays the role of 'cognition' that these other psychologists emphasized. Second, there is a strong primacy of time preferences Ainslie's theoretical account. The core of his model is a continuum of 'short-range interests' and 'long-range interests' on which lies the whole of self-control problems. This includes the most unconscious ones (e.g., itches and nail biting) to the most conscious ones (e.g., spending too much time building a career at the expense of leisure). From these two features, he has criticized dual models as being too focused on the cognitive aspects of self-control problems and/or (depending the dual model) as not considering the motivational aspects of long-range interests. He nevertheless considers that the 2012 extension of Fudenberg and Levine's dual model is the one that captures the most of his account (see Ainslie 2012). Finally, the sources of normativity in Ainslie's work are far too numerous and complex to be discussed here, but it should be noted that he clearly identifies exponential discounting and hyperbolic discounting as rational and irrational respectively, quite like economists and behavioral economists.

Summing up this subsection, we have seen that dual models' shared background from

see also Frantz (2005) and Egidi (2008) for discussions related to economics. For the criticism of the trend see esp. Keren and Yaacov Schul (2009; see also Keren 2013; Sahlin et al. 2010; Kruglanski 2013; Osman 2013); and for the answers by defenders, see esp. Evans and Stanovich (2013a;b); see also Monsell and Driver (2000) who make some similar points restricted to pure cognitive *P*sychology. On the normativity of logic and probability theory in the cognitive *P*sychology of reasoning and of judgment, see Stein (1996). On the trickier issue of the role of social norms in the debates within social cognition, see Bargh (1999).

*P*sychology is various, implying a variety of sources of normativity. Among these sources, some that are within the shared background are: dual modelers themselves (Schneider and Shiffrin), the unbiased anticipation of pleasure (Berridge) and the cognitive capacity to resist impulses (Shiv and Fedorikhin, Baumeister). Among those that are not in the shared background as characterized here are logic, probability theory and social norms (the dual trend) along with the motivational capacity to resist impulses (Ainslie). Notice also that, unlike dual modelers (with the exception of Bernheim and Rangel), all the psychologists reviewed here (with the exception of the critics to the dual trend) are pretty clear on the normative dimension of behaviors violating the aforementioned norms: they are *mistaken* or *irrational*. Finally, notice also that all dual modelers seek to articulate two domains of Psychology, pertaining to 'cognition' and 'motivation' respectively, that have long been separated. Though psychologists often debate the 'primacy' of motivation over cognition or *vice-versa*, this is not an issue with which dual modelers are concerned (with the exception of Loewenstein and O'Donoghue, especially in the 2004 early version of their dual model).

3.2.2 Thaler and Shefrin, Gul and Pesendorfer, and a neuroeconomics controversy

We now turn to the economics part of dual models' shared background, in much the same fashion as in the previous subsection. Three sets of references constitute this part of the shared background: the early work of Thaler and Hersh Shefrin in *behavioral economics*, the work of Gul and Pesendorfer in *standard economics*, and a controversy in *neuroeconomics*.

All dual modelers agree with Thaler and Shefrin's own claim to have provided "the first systematic, formal treatment of a two-self economic man" (1981, p.394). Constructed in two papers (1981; Shefrin and Thaler 1988), the core of their model is indeed extremely close to the ones discussed above. They model individual behaviors as the interactions between two types of self: one 'planner' optimizing a lifetime utility constituted by the aggregation of a succession of 'doers' optimizing the utility of each periods. The interaction is formalized in a principal-agent framework where the planner (i.e., the principal) can influence doers' (i.e., the agents') choices to avoid 'temptations' (i.e., consuming too much now at the expense of future doers). More precisely, the planner can influence doers *directly*, *externally* or *internally*,

but it represent utility costs that Thaler and Shefrin associate with the cost of self-control: 'willpower'. Direct influences require the most willpower, they consist in the planner changing the doers' preferences (intuitively, picture yourself trying to 'become another kind of human being' or 'improve your personality'). External influences require the least (but still some) willpower, they consist in (physical, social or institutional) precommitments, such as subscribing to saving plans and the like. These two influences correspond respectively to Strotz's strategies of consistent planning and of precommitment (cf. the previous chapter). Internal influences require an amount of willpower which is in-between the two previous types of influence, and consist in self-imposed rules, e.g., 'never buy new clothes more than twice a months'. Clearly, their two sub-individual entities are conceptually similar to Fudenberg and Levine's, and their interaction through direct influences is conceptually similar to both Fudenberg and Levine's and Loewenstein and O'Donoghue's. However, their interaction is *formally* closer to Brocas and Carillo's, but with limited control (of the planner over the doers) instead of informational asymmetry as the justification for the formalism. We shall see later that the interaction through internal influences that is at the core of Thaler and Shefrin's model is conceptually close to the applications of three dual modelers: Brocas and Carillo, Benhabib and Bisin, Fudenberg and Levine.[8]

Thus, the formalization of internal influences is the specificity of the core of Thaler and Shefrin's model with respect both to the literature that followed Strotz and to the cores of contemporary dual models. More precisely, in the core of their 1988 model, internal influences are formalized through '*mental accounts*', which refer to a development made by Thaler (1985) to his (1980) *micro*economic theory of consumer choice discussed in the first chapter of this dissertation. At the microeconomic level, Thaler argues that decision makers perform a mental arithmetic on the monetary consequences of their choices whereby mathematical operations are allowed between monetary consequences within a mental account but not between monetary consequences across mental accounts. But Shefrin and Thaler use a simplified version of mental accounting for macroeconomic analysis. To illustrate, consider "the following four events: a $1000 bonus received at work; a $1000 lottery windfall; a $1000 increase in the value of the

[8]While Bernheim and Rangel do not reference Thaler and Shefrin's model in the text of their paper, the latter is so pervasive in dual models that it still found its way into the former's bibliography!

household's home; and an inheritance, to be received in ten years, with a present value of $1000" (Shefrin and Thaler 1988, p.615). In any of these four events, do you consume goods you did not plan to consume (e.g., you treat yourself with a nice restaurant)? If yes, then you violate standard macroeconomic theory (permanent income or life-cycle hypotheses) that assumes consumption smoothing, i.e., that if one of these four events happens, you smooth your consumption through life (you consume the same thing you consumed but slightly more, you don't buy goods you did not planned buying). Furthermore, suppose the four events occur simultaneously, do you consider that you have an extra $4000 to spend or just $2000 (adding the first two, but not the last two)? If the latter, then it refines the explanation of why you violate consumption smoothing, namely because you do not treat money as being totally *fungible*. That is, you don't spend the same given amount of money in the same way depending on where the money comes from (e.g., inheritance *versus* gambling money), and on how large it is compared to the usual flow of money that comes from that source (e.g., a $1000 bonus on a $3000 salary *versus* on a $30000 salary). Shefrin and Thaler replace the subtleties of the microeconomics of mental accounting by only three broad mental accounts: "current spendable income (I) [where you probably put the bonus and lottery money], current assets (A) [where you probably put the home money], and future income (F) [where you probably put the inheritance money, but possibly some of the bonus and lottery money if your salary is around $1000]" (1988, pp.614-15, my comments in brackets). In the core of Shefrin and Thaler's model, internal influences of the planner on doers is formalized by assigning different marginal propensity to consume to each of these mental accounts, and the exact value of the marginal propensity is parametrized on a cost function representing the amount of willpower (which is thus account-dependent). The two main connections with contemporary dual models are these. First, Fudenberg and Levine (2006, p.1461; cf. next section) claim to provide microfoundations for Shefrin and Thaler's mental accounting, i.e., deriving the different marginal propensity from the core of their dual model instead of imposing them as assumptions. And in a recent extension for social preferences (Dreber et al. 2016; cf. next section), they have integrated mental accounting in the core of their dual model. Second, like Benhabib and Bisin, the core of Shefrin and Thaler's model can be taken as a contribution to microeconomics, but it is intended for applications to macroeconomics.

The normative dimension of the core of Thaler and Shefrin's model is, as with dual models,

an ambiguous issue; but by contrast with them (minus Bernheim and Rangel), it is an explicit issue. The ambiguity is exactly the one described in the introduction of this chapter. On the one hand, in their 1981 paper, they claim that the behaviors modeled are rational because self-control costs are an unavoidable constraint, and optimizing under that constraint is rational. But on the other hand, in their 1988 paper, they claim that rationality is to be equated with the standard theory and corresponds to the planner's preferences; the doers are irrational and the individual is *quasi-rational*. This shift is, I would argue, attributable to one thing that happened between the publications of these two papers, namely the beginning of Kahneman and Tversky's researches on *framing*. Indeed, Shefrin and Thaler explicitly present 'framing' with mental accounting as being the two new ingredients of their 1988 model from their 1981 one, the common ingredient to both being self-control. The role of framing in Shefrin and Thaler's model is deferred to the conclusion of this chapter. What is left to discuss here are the rather unusual sources of normativity in their model. In the 1981 model, human organizations, and especially firms are the main source of normativity. They argue that self-control behaviors are rational because they can be model, through the theoretical convention of the principal agent-framework, as following the same rules used by profit maximizing firms. In the 1988 model, the fungibility of money is main source of normativity. They argue that it is because decision makers treat money as not totally fungible that their behaviors are not totally rational (especially because it leads to framing effects, see the last section of this chapter). Indeed, as Thaler noted in 1985, when a system of mental accounting is formalized, "the *normative principle* of fungibility is relaxed" (Thaler 1985, p.201, my emphasis). In both cases and in the words of the entanglement thesis, value judgment of rationality or irrationality are made from both the factual observation of individual struggling with self-control problems and the theoretical conventions from economics.

While the work of Thaler and Shefrin can be seen as the behavioral economics region within dual models' shared background from economics, the work of Gul and Pesendorfer represents the standard economics region. Indeed, Gul and Pesendorfer have provided an axiomatic framework where self-control problems can be formalized with very minimal deviation from the standard model, and allowing various applications (2001; 2004a; b; 2005; 2007). The core of their contribution starts by changing the objects of choice over which preferences are defined. Instead of

being over the (probabilistic and/or temporally situated) consequences of a decision problem, Gul and Pesendorfer formalize preference over decision problems. To illustrate with a slightly modified version of their paradigmatic example (2001, pp.1402-3), consider the choice between two consequences: a vegetarian dish denoted by x and a hamburger denoted by y. In the morning (i.e., at $t = 1$), the decision maker can make a reservation for lunchtime (i.e., at $t = 2$) in a vegetarian restaurant where his choice set at lunchtime will be $\{x\}$, or in a hamburger restaurant without vegetarian dish hence with the choice set $\{y\}$, or in a restaurant where there will be both hence with the choice set $\{x, y\}$ that is the union of the two former, i.e., $\{x\} \cup \{y\}$. In other words, in the morning, the decision maker's preferences are over the set of decision problems $\{\{x\}, \{y\}, \{x, y\}\}$ that he can face at lunchtime. Drawing on the work of David Kreps, Gul and Pesendorfer observe that the standard model is characterized by the following implicit set-theoretic axiom: *in the morning*, a weak preference for the vegetarian dish over hamburger implies an indifference between, on the one hand, the vegetarian dish, and, on the other hand, the union of the vegetarian dish with the hamburger, i.e., $\{x\} \succsim \{y\} \Rightarrow \{x\} \backsim \{x\} \cup \{y\}$. Gul and Pesendorfer slightly strengthen this implicit axiom by interpreting a choice for the vegetarian dish in the morning as a precommitment against the self-control cost to resist the temptation of a simultaneously available hamburger, i.e., $\{x\} \succsim \{y\} \Rightarrow \{x\} \succsim \{x\} \cup \{y\} \succsim \{y\}$. They call this axiom *set betweeness*, and it is at the center of most of their formal results. They notably prove the equivalence of standard preferences respecting the independence axiom plus set-betweeness with, on the one hand, a utility functional with a self-control cost function, and, on the other hand, a choice function respecting the standard axioms of consumer choice theory. From this, Gul and Pesendorfer make three main contributions. First, they solve Strotz's strategy of consistent planning without using dynamic game theory i.e., without implying multiple selves. Second, they characterize the revealed preference implications of behavioral economics' models with quasi hyperbolic discounting. Third, they formally represent intertemporal preference reversals in a dynamically consistent model. Their applications are in macroeconomics – related notably to the work of Laibson, Thaler and Shefrin and Benhabib and Bisin – and in the economics of addiction – related notably to the work of Bernheim and Rangel.

In terms of the sources of normativity, Gul and Pesendorfer explicitly endorses the standard view depicted in the two previous chapters. That is, on the one hand (cf. chap.1), they consider

that the behaviors captured by their model are rational in so far as they respect consistency conditions on a choice function which implies the optimization of a utility function (see all the references cited in the previous paragraph). And on the other hand (cf. chap.2), they judge the independence axiom, which is always part of their axiomatic framework, as normatively justified because of its intuitive appeal (see Gul and Pesendorfer 2008, p.29). Gul and Pesendorfer's emphasis on the rational interpretation of the behaviors plays a specific role with respect to Fudenberg and Levine's dual model. To understand how, we need to briefly state one feature of the latter's dual model which is in-between its cores and its applications, namely its possibility to represent self-control costs that are convex (i.e., non-linear). Roughly, the issue of linear *versus* convex self-control cost function is about whether the marginal cost of self-control is proportional (linear) or more than proportional (convex) to the forgone marginal utility it entails. Fudenberg and Levine (2006, section VI) compare their *modeling approach* to self-control with Gul and Pesendorfer's type of *axiomatic approach*. Fudenberg and Levine show that with linear costs of self-control, the preferences induced by their dual model are consistent with the axioms of Gul and Pesendorfer (2001). But this is not the case anymore when self-control costs are not linear, as their dual model then violates some of the most basic axioms: contraction consistency on choice functions (cf. chap. 1), and the independence axiom on preference relations (cf. chap.2). The point I wish to make here is twofold. On the one hand, Fudenberg and Levine do not see the violation of the standard axioms as necessarily entailing irrationality, by contrast with the standard view. They rather argue that if convex self-control is an internal constraint against which the decision maker cannot do anything, then "insights from psychology and neuroscience" becomes useful to characterize what counts as rational (see Fudenberg and Levine 2006, p.1469). On the other hand, their work have motivated further contributions in the axiomatic approach (surveyed by Lipman and Pesendorfer 2013, see esp. sect.7), which have themselves motivated further extensions of Fudenberg and Levine's dual model (2012c). In the words of the entanglement thesis, value judgment of rationality or irrationality depends on the factual observation of dynamically inconsistent behaviors and the theoretical conventions of economics (for both Gul and Pesendorfer and Fudenberg and Levine) and *P*sychology (for Fudenberg and Levine only).

The last part of dual models' shared background from economics is a controversy in neuroe-

conomics that can be circumscribed to two sets of papers. On the one hand, two papers involving behavioral economists Laisbon and Loewenstein (McClure et al. 2004; 2007), and on the other hand, one paper involving neurobiologist Paul Glimcher (Glimcher et al. 2007). Both presented decision maker with the type of intertemporal choice discussed in the previous chapter while scanning their brain activity. In terms of behavioral results, McClure et al.'s were consistent with β-δ preferences (i.e., quasi hyperbolic discounting), and Glimcher et al.'s with hyperbolic but not quasi hyperbolic discounting (i.e., the sooner-smaller consequences were chosen whether they occurred immediately or with some delay). In terms of neural results, Glimcher et al. use their observations to argue for a unitary neural system of valuation, contradicting multiple selves and dual models of discounting. By contrast, McClure et al. use their observations to argue for a neural foundations of quasi hyperbolic discounting: a 'β-system' strongly activated when immediate consequence are available (related to the mesolimbic dopamine system), and a 'δ-system' activated continuously regardless of the timing of consequences (unrelated to the mesolimbic dopamine system). We shall not go into the detail of this opposition, but only note two different ways of using neuroeconomics references by dual modelers, who cite both references but emphasize on McClure et al's. One way, taken by Fudenberg and Levine and Loewenstein and O'Donoghue, is to argue that dual modeler's two sub-individual entities are *not inconsistent* with what is known about the brain. The other way, taken by Brocas and Carillo, is to take McClure et al's findings as the neural foundations from which theoretical models of individual behavior have to be constructed. In short, neuroeconomics is definitely part of dual models' shared background, but it is clearly central only for Brocas and Carillo.[9]

The two main sources of normativity in neuroeconomics are not traditional for standard economics. One source is medical. Behaviors associated with the activation of brain areas that are similar to the activations known to be characteristic of pathological subjects (e.g.,

[9]It should be note that this is only the intertemporal part of a larger controversy which is constitutive of the 2000s development in neuroeconomics. On the one hand, there is a trend constituted mainly of behavioral economists seeking neural evidence for behavioral economics models' and against standard economics' models; on the other hand, there is a trend constituted mainly of neurobiologists using the axiomatic framework of standard economics to theorize about the behavior of the brain's neural activity. The issue is mainly about the use of formal language, especially about the helpfulness of the axiomatic method from economics for the computational ones used in *P*sychology and the neurosciences. The specificity of the intertemporal part of this controversy is that no camp claims any positive adequacy of standard exponential discounting. (see, e.g., Rustichini 2008; Glimcher and Fehr 2014; or Vromen 2011; Vallois 2012b; Ross 2014 for a reflexive perspective). Also, notice that for chronological reasons both Bernheim and Rangel and Benhabib and Bisin don't cite McClure et al.'s and Glimcher et al.'s references. Arguably, the latter would side with Brocas and Carillo's way, while the former would be in-between them and the other dual modelers.

schizophrenic, sociopath, drug addicts etc.) are deemed to be irrational. The other source is the distinction between Man and animals. Behaviors associated with high activation of brain areas that are known not to be shared with animals are deemed rational. The former source is explicitly used by all dual modelers with the exception of Fudenberg and Levine, while the latter is explicitly used by Loewenstein and O'Donoghue only. Again, it is beyond the scope of this dissertation to go into the details of these biological sources of normativity (see references in the preceding footnote).

Summing up this subsection, dual models' shared background from economics is less various than the one from *P*sychology, yet it contains various sources of normativity. Among these sources, there is, of course, the traditional one from consistency and optimization arguments (Gul and Pesendorfer, and to a certain extent, the 1981 version of Thaler and Shefrin where self-control costs are an inevitable constraint under which to optimize). But there are some others which are far less traditional: the analogy of self-imposed rules with profit-maximizing firms (Thaler and Shefrin 1981), the fungibility of money (Shefrin and Thaler 1988), and the two biologically driven distinctions (neuroeconomics): the medical pathological/normal one and the human/animal one. We have pointed one source of normativity related to 'framing' but have left it undiscussed until the conclusion of this chapter. Finally, given the role played by Thaler in the constitution of behavioral economics and the role of neuroeconomics with Gul and Pesendorfer (2008) in the 2000s 'behavioral *versus* standard economics debates', this subsection suggests that dual models are indeed at the center of the reconciliation between standard and behavioral economics.

Conclusion

We can summarize the two subsections of this section very briefly regarding the purpose of this chapter. On the one hand, the psychologists studied in the first subsection, though drawing on a variety of sources of normativity, invariably and implicitly favor an interpretation where the sub-individual entity that evaluated as optimal to take neither cake nor coffee in the preceding section is rational, and any deviation from it cannot be normatively justified. On the other hand, the standard economists studied in the second subsection arguably favor the alternative implicit interpretation where the decision maker's evaluation (i.e., just coffee is optimal) is rational, so

that a normative interpretation can be given to the core of a dual model, but not to any of its underlying sub-individual entities (one is 'too' rational, while the other is irrational). However, the role of 'framing' from behavioral economics in that latter interpretation may complicate matters a little.

3.3 Applications: theoretical unification through self-control

This section pursues the comparative studies of dual models by focusing on their applications. We briefly discuss various applications that are peripheric to the empirical and theoretical issues discussed in the previous part of this dissertation (3.3.1). We then turn to applications directly related to these issues, by highlighting a primacy of time preferences (3.3.2) and a primacy of risk over social preferences (3.3.2) for dual modelers.

3.3.1 Applications peripheric to the primacies of time, risk and social preferences

Macroeconomics and the economics of addiction are the two main domains of application that lie outside of last part's empirical and theoretical scopes. They are respectively associated with Benhabib and Bisin's and Bernheim and Rangel's dual models. Here we shall briefly illustrate how their applications proceed and note the implications that are relevant from the methodological perspective of this dissertation.

Benhabib and Bisin are interested in a representative agent's consumption-saving rule*s* within a dynamic economy with stochastic temptations (unplanned opportunity to consume for immediate utility). Only automatic processes derive utility from stochastic temptations, determining a certain marginal propensity to consume different from the marginal propensity to consume of controlled processes. The theoretical unification achieved by Benhabib and Bisin stems mainly from the possibility of their dual model to generate different consumption-saving rules, i.e., with the automatic *or* controlled processes' marginal propensity to consume. The representative agent adopts either of these rules under different circumstances, depending on the size of stochastic temptation and of the attention parameter. This allows them to account for empirical regularities that models from neither standard (i.e., with life-cycle and permanent income

hypotheses) nor behavioral (e.g., Laibson's work) macroeconomics can account for. The clearest example of such regularity is the observation of different marginal propensity to consume out of income shocks (akin to stochastic temptations) *depending* on the size of the shock (the bigger it is, the smaller the propensity to consume; see Benhabib and Bisin 2005, pp.478-9).

Bernheim and Rangel are interested in the conditions under which drug addicts make *mistakes* in their choices to consume drugs. Making mistakes consists in planning not to consume drugs in the cold mode but then entering in the hot mode and consuming drugs. The theoretical unification achieved by Bernheim and Rangel stems mainly from the possibility of their dual model to generate such mistakes without assuming that the decision maker has conflicting preferences (and he can furthermore rationally anticipate the risk of such mistakes in the cold mode). The clearest example of an empirical regularity that neither standard nor behavioral economics' models of addiction can account for is what they call 'self-described mistakes': choices of drug consumption that are described as a mistake by the addict *while he is consuming the drug*. That is, his current choice goes against his current preferences. Standard models cannot account for it because of their methodological commitment to revealed preferences (esp. in Gul and Pesendorfer's work). Behavioral economists' models rely on multiple selves and cannot account for it because for the self who is consuming, choice matches preferences. It is only from the perspective of another (past or future) self that the choice can said to be mistaken (esp. in Laibson's work and his followers). But then the mistake is not recognizing *during the act of consumption*.

More generally, the main similarity between the applications from these two couples of dual modelers is that it can be deemed to be a theoretical exercise for the following reasons. Their applications consist in conducting intertemporal comparative statics (i.e., comparative dynamics) to derive the economic implications of the cores of the their dual models. Both emphasize that the economic behaviors thus generated account for empirical regularities that standard and behavioral economics models can *and cannot* account for. These empirical regularities are only used as *indirect evidence*, i.e., qualitative patterns the model would predict, but none of them calibrate their dual models with existing data to predict other existing data. Finally, both emphasize their achievement of theoretical unification by pointing out how variations of parameter values can reduce their dual models to either standard or behavioral existing ones.

For Bernheim and Rangel, when the hot mode probability distribution is set to zero their dual model reduces to standard models (where consumption of temptations is always the result of a utility calculus in a cold mode style). And they explain what modifications would allow a multiple selves reinterpretation of their models, which they criticize (see 2004, pp.1581-2). For Benhabib and Bisin, theoretical unification is more straightforward: when the attention cost parameter is set to zero, then it reduces to standard models (where temptations are avoided at no costs), and when it is set to infinity, then it reduces to behavioral models (where temptations cannot be avoided without external precommitments).[10]

Not much can be directly learned about the positive/normative issue w{ithin dual models from these applications. However, the role played Bernheim and Rangel's in *The Foundations of Positive and Normative Economics* (Caplin and Schotter 2008a) is instructive. Roughly, it lies in Gul and Pesendorfer's (2008) criticisms of their modeling of 'cues' and 'mistakes' within Bernheim and Rangel's dual model. 'Cues' was indeed one of the central term at stake in the Gul and Pesendorfer (2008) *versus* Camerer (2008) exchange framed as an issue over the use of formal and ordinary language by economists (cf. the general introduction). Gul and Pesendorfer's criticism of 'mistakes' triggered most of the discussion of the positive/normative issue in *The Foundations*. They targeted the only everyday life example (i.e., not about addiction as theorized by scientists) offered by Bernheim and Rangel (2004, pp.1561-2). This example is about the following empirical regularity: American tourists crossing streets in Great Britain often get injured because they tend to look only to the left before crossing, though cars arrive from the right. Bernheim and Rangel's point is that choice of looking only to the left before crossing would be better modeled as a systematic mistake rather than as a revealed preference for doing so. Gul and Pesendorfer (2008, pp.22-23) disagree and argue that a revealed preference approach is possible even in such cases. Mistakes can be modeled as "subjective constraints on the feasible strategies that are not apparent from the description of the decision problem" (ibid), and that such constraints should be motivated by observable data such as "data showing that American tourists in London avoid unregulated intersections" or "that a sign that alerts the American tourist to "look right" alters the decision even though such a sign does not change the

[10] Brocas and Carillo (2008, sect. III.B) venture into a brief application of their dual model to macroeconomics, where they note that they cannot explain the strong correlation between unexpected income changes and marginal propensity to consume that Benhabib and Bisin are able to explain.

set of alternatives" (ibid). Such a revealed preference approach is in fact close to what Bernheim and Rangel propose in their normative contributions that are not restricted to the economics of addiction (see 2007a; b; 2008; 2009; Bernheim 2009). Their only difference with Gul and Pesendorfer is that they prone an interdisciplinary approach where choice data are not anymore the only type of data that would help to characterize mistakes; their use of neurosciences in their dual model is always taken as an example of that, as already noted in the first section above. Benhabib and Bisin (2008) have the exact same position on the issue of mistakes (see esp. pp.325-6). Furthermore, in their discussion of mistakes, though they remain implicit about the normative dimension of their own dual model, it is obvious for them (as it is intuitive as well) that mistakes are the produce of automatic processes (see esp. p.327).[11]

Summing up in one sentence, although Bernheim and Rangel and Benhabib and Bisin's dual models do not tackle the empirical and theoretical issues at the center of the standard *versus* behavioral economics debate, they (esp. Bernheim and Rangel) played a non-trivial role in the underlying methodological issues of this debate.

3.3.2 The primacy of time preferences in dual models

Having discussed Bernheim and Rangel's and Benhabib and Bisin's applications, we let them aside to concentrate on the remaining three couples of dual modelers. The simplest application to time preferences is Loewenstein and O'Donoghue's in the published version of their dual model with Bhatia (Loewenstein et al. 2015, pp.65-61). The deliberative and affective system's utility functions are both endowed with exponential discount functions, but the deliberative system discounts at a lower rate than the affective system. Furthermore, the latter's discount function depends on the intensity of affective motivation parameter: the higher this parameter, the lower the discounting from the affective system. Theoretical unification for time preference is achieved straightforwardly by Loewenstein, O'Donoghue and Bhatia. They show that the previous characterization yield a utility function *for the decision maker* with a discount func-

[11] It should be noted that Gul and Pesendorfer (2008, pp.10-13) mistakenly attribute *some* quotes from Bernheim and Rangel (2004) to Camerer, Loewenstein and Prelec (2005). We shall not discuss the issue of 'mistakes' in *The Foundations* further because it belongs to the positive/normative issue at the level of subareas in economics (see esp. Bernheim and Rangel's notion of 'ancillary conditions' and 'multiself Pareto optimality' in the references mentioned above ; see also Köszegi and Rabin 2008a;b). Note also that how Camerer's (2008, pp.55-56) take on the issue of mistake and feasible constraint is even more illustrative of the role of not only technical ordinary but also everyday ordinary language in the interdisciplinarity issue (cf. the general introduction).

tion that does not have a constant discount rate, but one that varies according to the three parameters of their dual models. Discounting gets closer to impatience and even impulsivity if cognitive load increases, and/or willpower strength decreases, and/or the intensity of affective motivation increases. These parametric variations allow them to derive smooth hyperbolic discounting for the decision maker (i.e., à la Ainslie), as well as quasi hyperbolic discounting (i.e., à la Laibson) in the special case where the intensity of affective motivation is so great as to imply a discount rate that tends to infinity for the affective system. In that latter case, the constant β from quasi hyperbolic discounting becomes a variable which depends on the cognitive load and willpower strength parameters. In short, because their dual model nests both hyperbolic and quasi hyperbolic discounting (and also exponential discounting with appropriate parametrization plus assuming the same optima for both systems), it achieves theoretical unification for economics; and the threefold parametric variations allow to achieve theoretical unification of finer phenomena that are more of interest to psychologists working on ego depletion.[12]

Brocas and Carillo's (2008, sect. II) main application to time preferences also starts by distinguishing their two sub-individual entities by different time horizons: long-term for the principal brain system, short-term for the agent brain system. Their choice set is constituted by a sequence of triples occurring at each time periods: one desirable action, one undesirable action, and an information parameter about the relative desirability of both actions. On the one hand, the principal can control the agents' actions without utility costs and derives utility from the undiscounted sum of agents' utilities. But on the other hand, only agents have access to the information parameter, and, contrary to the principal's optimum plan, agents optimize at each period regardless of the future. The interaction between both, i.e., how the former elicit information from the latter and design contracts for them, depends on an intertemporal constraint (akin to a budget constraint that has to be met at the last period) containing the information parameter. Their theoretical unification is achieved less straightforwardly than Loewenstein et al. (2015). First, they formalize a two-periods consumption-labor problem (with borrowing and saving) within this setup. In this problem, the second best optimal contract designed by the principal can be interpreted in ordinary language as reflecting the kind of "self-

[12]Loewenstein and O'Donoghue's (2004; 2007) applications to time preference have been subjected to great changes across versions. Notably, they used to separate their accounts of intertemporal choice from an account of temptation and claimed that their dual model also nests Gul and Pesendorfer's models.

disciplining rule" (i.e., internal influence à la Thaler and Shefrin) we usually impose on ourselves in everyday life. Brocas and Carrillo put it as ""I will go to this dinner party only if I first exercise for an hour"" (2008, p.1322). Second, they extend the two-periods setting to an indefinite (but not infinite) horizon. In this extension, the previous asymmetric information case generalizes into a discount function. Because they did not assume constant discounting, the discount rate in their function is endogenous in the sense of being a variable depending on the characteristic of the informational asymmetry. They stress how this function can exhibit the property of hyperbolic discounting and is problem-dependent, accounting for the diversity of discount rates mentioned in the previous part. Third, they slightly modify the consumption-labor model by postulating that the two actions are either two consumption actions (e.g., food and clothing) or two labor actions (e.g., working in the morning or working in the afternoon), and argues that this captures mental accounting phenomena. Finally, Brocas and Carillo (2008, sect.IV) modify the core of the model depicted so far by abstracting from its temporal dimension. They do so by drawing on Berridge's work explained above: the principal system optimizes the liking of pleasure and derives utility from the agent system's utility, which however optimizes the wanting of pleasure. The bias between liking and wanting is again captured by an information parameter, and the formal second best contract by the principal can be interpreted in ordinary language as reflecting internal influences à la Thaler and Shefrin, e.g., ""as long as you don't abuse, you can do whatever you want "" (2008, p.1330). This result can be seen either as the special case of the previous model for one shot decision, i.e., its very core, or as another dual model built from their methodology.

We can start by illustrating one application of Fudenberg and Levine's to time preferences with linear cost of self-control (2006, sect.II), i.e., the cost in utility to avoid a given amount of consumption is proportional to the marginal utility of that amount. In an infinite horizon consumption-saving (without borrowing) setting, short-run selves' actions are assumed to determine the saving rates, which would be zero so that all wealth is consumed by the first short-run self without the intervention of the long-run self. In this case, the optimal saving rate is a *constant* part of lifetime wealth but the size of that part varies depending on the ratio of proportionality underlying the linearity of self-control costs. In the case with short*er*-run selves, i.e., living more than one period, Fudenberg and Levine refine this result showing how it can

also depend on the length of time over which a shorter-run self derives utility (2012b, pp.15-13). These two results nest quasi-hyperbolic and hyperbolic discounting and allow for parametrization generating the diversity of discount rates mentioned in the previous part. But the greatest source of unifying power in Fudenberg and Levine's dual model stems from its possibility to generate convex (i.e., nonlinear) costs of self-control, i.e., the cost in utility to avoid a given amount of consumption is superior to the marginal utility of that amount. It is the experimental results of Shiv and Fedorikhin (1999) on the positive relation between cognitive load and the cost of self-control that first motivated Fudenberg and Levine (2006, sect.V) to explore this formal possibility of their dual model. To do so, they simply provide a convex specification of the self-control cost function with a cognitive load variable that increases the marginal cost of self-control for a given consumption, i.e., for one such consumption, the higher the cognitive load the higher the utility lost by the long-run self to control the short-run self. This account of cognitive load is static. Fudenberg and Levine (2012c) made it dynamic by drawing on the work of Baumeister and replacing the cognitive load variable with a stock of willpower. They formalize a depletion-repletion dynamics whereby willpower resources not used in one period carry over in the next one. Roughly, by exploring combination of linearity and nonlinearity of the stock of willpower's (1) depletion, (2) repletion and (3) impact on the cost of self-control, Fudenberg and Levine (2012c, sects. 5-6) provide a formal account of ego depletion. In the previous section, one reason why convex cost of self-control loom large in Fudenberg and Levine's work was already discussed: it violates the independence axiom and the condition of contraction consistency on choice function, but Fudenberg and Levine (2006, sect.VI) do not necessarily see these violations as irrational. Another reason is that the violations of the independence axiom are worked out by Fudenberg and Levine so as to account for Allais' paradoxes for risk preferences and the interactions of risk and time preferences. Only one of these interactions fits this subsection, namely the introduction of risk in the dimension of time. The intuition in how they model the latter is quite simple: the marginal cost of self-control not only depends on the timing of the consequence, i.e., the sooner the bigger the cost, but also on its probability of occurrence, i.e., the surer the bigger the cost. Fudenberg and Levine (2011, pp.43-44) show that when both of these functional relations are convex, there dual model implies the qualitative pattern of Keren

and Roelofsma's (1995, exp.1) data discussed in the previous chapter.[13]

We can sum up the applications of these three dual models to time preference by pointing that all achieve theoretical unification in the same way. Namely, they turn the standardly *constant* discount rate into a *variable* which depends on further parameters reflecting the *p*-&*P*sychology of self-control. The unification is theoretical in the sense that it nests existing (hyperbolic and quasi-hyperbolic) models already accounting for the variability of discount rates (cf. previous chapter). And it is interdisciplinary in the sense that it formally accounts for ego depletion phenomena dear to psychologists. Finally, notice that there is a primacy of time preference in each dual model the following senses. Brocas and Carillo (2008) model *only* behaviors that are traditionally discussed in relation to time preferences. Therefore, in the next subsection they will be put aside, so that there will remain only two couples of dual modelers left. While Loewenstein and O'Donoghue model behaviors over time, under risk and regarding other people *separately* (in a different section for each, as we shall see), they nevertheless state that "[t]he most straightforward application of our model is to intertemporal choices" (Loewenstein et al. 2015, p.16). Fudenberg and Levine start by modeling the classical economic application of intertemporal choice, namely consumption-saving behaviors, before modeling the less classical ones involving cognitive load and other ones related to risk and social preferences. And as we shall see in the next subsection, they necessarily introduce a temporal dimension in their modeling of behaviors that are classically considered as being uniquely under risk or uniquely regarding other people.

What can be learned about the positive/normative issue within dual models from the applications discussed here? I propose to answer this question only for Brocas and Carillo here, and defer the answer for the two other couples of dual modelers to the next subsection (once all their applications would have been discussed). In lights of their applications discussed here, a set of earlier methodological manifestos from Brocas and Carillo allow to clarify where the ambiguity on the positive/normative issue comes from in their work (2003a; b; 2004a; Brocas et al. 2004). Put roughly, they share the same ambiguities already discussed in this dissertation: the two-ways direction of influence between the two levels of the normative/positive issue (see

[13] Another application to time preference (with linear self-control cost) include procrastination (2006, sect.IV) where they argue that their model nests Rabin and O'Donoghue's model (cf. previous chapter) with more analytical tractability and better predictions.

e.g., 2003b, p.90), taking standard economics and the economic agents as sources of normativity while ignoring possible conflicts between both (see e.g., 2003a, p.xiii). With respect to Brocas and Carillo's dual model, the latter problem implies either to count as irrational the self-disciplining rules we use in everyday life, or to count them as rational which implies seeing their dual model as a new standard economics where our everyday quasi-rational behaviors are the new norms. An aspect of Brocas and Carillo's position complicates the picture and is highly specific to their background as economists (i.e., contract theory and mechanism design), namely their reliance on strategic *ignorance* as a potential source of normativity. Indeed, their applications of their dual model suggest that it may be optimal not to fully know oneself's preferences, i.e., to ignore a little bit what one's tastes are (as the saying goes, 'ignorance is bliss'). Brocas and Carrillo (2003b, sect.3.3; Brocas et al. 2004 sect.3) make it clear that this is the translation to intrapersonal conflict of well established results about the efficiency of commitments to ignore information in interpersonal conflicts. Although counting everyday quasi-rational behaviors as the norm to be enforced can be seen as an interesting perspective, doing that by counting Brocas and Carillo's dual model as the new normative benchmark where the main source of normativity is ignorance faces non-trivial objections. Generally speaking, Sen's (1987, p.86) comments on the use of strategic ignorance at the interpersonal level might still applies to their intrapersonal translation: "[i]t may be that people are often ignorant, but a model of 'rational' behaviour that counts on ignorance for being able to achieve good results, which will fail to be realized if people become better informed, has an element of perversity in it" (see also Walsh 1996, p.200). [14]

3.3.3 The primacy of risk over social preferences in dual models

One of the typical behavioral phenomena under risk that Fudenberg and Levine (2006, sect. III) account for through their dual model (with *linear* self-control costs) is the Rabin paradox. Recall that Rabin (2000) showed that plausible risk aversion over win/lose small stakes implies implausible risk aversion over large ones with expected utility theory (c.f., previous chapter). Fudenberg and Levine's account of the Rabin paradox starts from a modification to their simple consumption-savings model: each period is divided into two subperiods, a "bank" subperiod

[14] For a radical criticism of the normative implications of the use of 'ignorance' in standard economics, see Mirowski and Edward Nik-Khah's forthcoming book.

and a "nightclub" subperiod. In the bank subperiod, consumption is impossible so that the long-run self decides *without self-control costs* how much to save and thus how much "pocket cash" is carried in the nightclub subperiod. In the latter, a short-run self consumes out of pocket cash and the rest goes to the next period's bank. But in each nightclub subperiod there is a probability which cannot be anticipated by the long-run self that lotteries are among the possible consumption available from pocket cash. If that probability does not occur, then *again* the long-run self does not need to impose self-control cost so that the short-run self consumes all the pocket cash. If that probability occurs, then the long-run self will use self-control to influence the short-run self's evaluation of the lotteries. Fudenberg and Levine (2006, Theorem 2) determine the conditions under which a lottery will be evaluated *only* regarding consumption out of pocket cash, or *also* regarding consumption out of wealth (i.e., future periods' spendings and savings). Roughly, these conditions ultimately bear on the size of the expected consequence from the lottery. And the long-run self only pays self-control costs if that size is big enough so that integrating (at least a part of) it into lifetime wealth has a marginal utility superior to these utility costs. This double standard of evaluation depending on the size of lotteries' consequences implies the Rabin paradox: small stakes are evaluated with respect to pocket cash and large stakes are evaluated with respect to lifetime utility so that the former evaluation is made through a higher level of risk aversion than the latter. Notice two features of Levine and Fudenberg's (2006) account of the Rabin paradox: (1) only the size of consequences plays a theoretical role, not the size of the probabilities, and (2) self-control costs are linear. As already glanced in the previous subsection, they later allowed both the size of probability to play a theoretical role in determining short-run selves' temptations *and* convex costs of self-control within the previous setup to account for further phenomena related to risk preferences. Among them are the Allais paradoxes and the pattern of preferences due to the introduction of time in the dimension of risk discussed in the previous chapter (see Fudenberg and Levine 2011, pp.45-48).[15]

Social preferences are currently being modeled within Fudenberg and Levine's dual model in a working paper with economist Anna Dreber and psychologist David Rand (Dreber et al.

[15]Because the explanation of the Rabin paradox does not need convex self-control cost, Fudenberg and Levine (2006, p.1461) discuss how quasi-hyperbolic discounting in a proper setting could also explain the Rabin paradox. There is also further phenomena related to risk preferences that Fudenberg and Levine account for but that were not discussed previously in this dissertation (see Fudenberg and Levine 2011, pp.48-49; see also the simplification of the dual model for risk preferences by Fudenberg et al. 2014).

2016). They introduce altruism as "a positive concern for others", i.e., a new "worthiness" parameter in Fudenberg and Levine's dual model with short*er* run selves (2012b; c) and for one shot-decisions (following Fudenberg et al. 2014's simplification). The setting is indeed constructed for simple binary allocations of money where utility is derived from money spent on consumption for "*me*" (i.e., the decision maker), but also from the money "*you*" (i.e., the other) receives thanks to *me*. It is on this latter source that the worthiness parameter applies: the smaller this parameter, the smaller the utility derived from a given amount of money received by *you*. There are three periods instead of *sub*periods, and everything that happens after the third one is 'the future' from which only the long-run self derives utility. In the first 'mental accounting' period, without any cost of self-control (because there is no consumption in the near future), the long-run self sets a "spending limit" constraining the second period. In the second "decision period", money from two sources has to be allocated between *me* and *you*: from found money that was unexpected in the previous period, on the one hand, and from lifetime wealth, on the other hand. Dreber et al. make two benchmark assumptions. First, that the long-run and shorter-run selves have identical concerns for *me* and for *you*, i.e., the conflict is only over time, not regarding people. Secondly, that in the absence of found money, it is optimal to give nothing to *you* and to spend only the predetermined spending limit for *me*. In the third "consumption period", the utility from spending the allocation on *me*'s consumption and for *you* is received. What matters about this last period is the time between it and the previous one, determining the lifespan of a given shorter-run self. Dreber et al. (2016, Theorems 1-2) determine the parametric conditions under which it is optimal to give money to the other only from unexpectedly found money, and how, in this case, the optimal amount to be given depends on the amount found by *me* and received by *you* (the latter includes the possibility of a transfer multiplier, e.g., if *me* gives €5, *you* receives €10). Among the six phenomena that Dreber et al. seek to explain, only two were discussed in the previous chapter. The first one is the donations in dictator games that can be (1) all, (2) nothing, (3) half or (4) in-between amounts of a given sum. Roughly, the explanation here comes from their assumptions (a) that this sum was not anticipated in the first period by the long-run self and (b) that the worthiness parameter is neutral (i.e., $= 1$). In this case their formal results (Theorem 2) can predict all of the four cases through variations of the other two parameters (self-control costs and shorter run self's

discounting). The second phenomenon is the introduction of time in the dimension of other people displayed in Kovarik's (2009) experiments. Roughly, the explanation relies on the effect of the time between the decision and consumption periods on the previous results. Because the longer this time the bigger the discounting, it increases the thresholds above which the decision maker gives everything or something (hence reducing average sharing in the future for a given amount of money). Notice that convex costs of self-control play no role in the explanation of these two phenomena, though it does for others not discussed in the previous chapter.[16]

Loewenstein, O'Donoghue and Bhatia's (2015) theoretical unification for risk (pp.65-71) and social (pp.71-73) preferences are more straightforward, and follow roughly the same logic of their theoretical unification for time preferences discussed in the previous section. We can therefore discuss both in parallel instead of separately. The first two modeling steps that are common to the three of Loewenstein et al.'s theoretical unification are, firstly, to change the initial choice set to a set of lotteries for risk preferences and a set of binary allocations between the decision maker and another person for social preferences, and, secondly, to specify the two systems' utility functions. For risk preferences, they assume that the deliberative system's utility function is expected utility for non-monetary decisions and expected value for monetary decisions; and that the affective system weights probability non-linearly and exhibit loss-aversion in the evaluation of consequences. For social preferences, they focus on altruism by assuming that the deliberative system puts a stable weight on the utility he derives from the consequences to the other while the affective system puts a variable weight on them, this weight being just the intensity of affective motivation parameter. In both cases, the interaction of the two systems' utility functions is mediated by the cost of mobilizing willpower in a way that can steer the individual utility function to account for phenomena or results that are usually accounted for by models from *either* standard *or* behavioral economics. Yet the theoretical unification achieved for risk and social preferences can be methodologically distinguished as follows. For risk preferences, the theoretical unification is theoretical in roughly the same sense as theirs for time preferences. They show how their dual model of risk preferences nests standard and behavioral economics'

[16] Briefly, the other phenomena are the amount people tend to give to beggars on the street which has nothing to do with the amount they tend to give to other anonymous in the lab, the effects of cognitive load and time pressure on decision maker's allocation (where convex costs of self-control play the main explanatory role), and the tendency to "avoid the ask" (e.g., crossing the street not to give to a beggars we would have otherwise given to). Notice also that Dreber et al. (2016) is the paper mentioned in the previous section where Fudenberg and Levine makes reference to the dual trend in Psychology for the first time.

models through parametric variations. So it can explain the regularities already explained by the latter. They notably discuss the Rabin paradox and the endowment effect under certainty, both explained principally through the affective system's loss aversion. They indeed emphasize that their dual model can easily approximate Kahneman and Tversky's original (1979) prospect theory, while making further predictions concerning the effect of ego depletion on risk taking. For social preferences, the theoretical unification cannot be achieved in the same sense as for the two other types of preferences, because there is no clearly identified standard *versus* behavioral economics' models (see the plurality of models in the previous chapter; and Loewenstein and O'Donoghue 2004, p.34). In principle, their account of social preferences allows them to formally account for any possible binary allocations. But it would be done in a very narrow data-fitting way by changing the intensity of affective motivation parameter for each change in allocations. They indeed explicitly acknowledge (p.73) that more structure would be needed to apply their dual model to the motivations underlying the plurality of results in the dictator game, which are not all affective (e.g., equality, fairness and the like). Again, they focus more on the effect of ego depletion or pure affective features (e.g., showing pictures instead of numbers in charity requests) of decisions related to social preferences.

The applications of Fudenberg and Levine's and Loewenstein and O'Donoghue's dual models discussed in this and the previous subsection can be contrasted in two respects: their relations to empiricism and the positive/normative issue. Their relation to empiricism nicely illustrate these dual modelers' backgrounds as economist. On the one hand, Loewenstein and O'Donoghue's backgrounds as behavioral economist is illustrated by the great number of empirical evidence they cite from neurosciences and animal studies to justify the assumptions they use for each system's utility function. For each of the three types of preferences, they systematically make two new predictions (one from the cognitive load and willpower parameters, the other from the intensity of affective motivation parameter) for which they encourage the conduct of experiments. Concerning the very empirical phenomena they seek to give a unifying explanation of, however, Loewenstein and O'Donoghue (and Bhatia) usually satisfy themselves with showing that their dual model nests the corresponding behavioral economics model that already explain them, and do a little comparative statics (or dynamics) to show that their predictions go qualitatively in the right direction. On the other hand, Fudenberg and Levine's backgrounds as standard economist

is illustrated by the quasi absence of evidence used to motivate the behavioral assumptions: it comes more from introspection and casual empiricism for them than it does for Loewenstein and O'Donoghue, maybe with the exception of their uses of Shiv and Fedorikhin's (1999) results to motivate convex costs of self-control. The new predictions they derived from their dual model are much less systematic than Loewenstein and O'Donoghue, though they do make some non-trivial ones, such as the one about the interaction between time and risk (see 2006, pp.1461-2) or social (see ibid, p.1458) preferences. And the comparative dynamics they conduct is more thorough, using different functional forms in appendixes to encourage standard economists to use their dual models, with some non-trivial success, notably Andersen et al.'s (2008) econometric paper discussed in the previous chapter. The behavioral *versus* standard contrast can be illustrated further by pointing out that Fudenberg and Levine's dual model would have a hard time to work with experimental data on hypothetical choices about losses, as the decision maker can only lose money he does have. But the contrast is also severely blurred by pointing out that their 2016 paper is co-written with a psychologist (Rand) and an experimental economist (Dreber), notably because they wanted more data on the interaction between time and other people than Kovarik's (2009), and conducted their own experiments accordingly.[17]

Finally, it is easier to discuss the positive/normative issue by treating time with risk preferences separately from social preferences. In Fudenberg and Levine's applications, most of the challenges from behavioral economics about risk and time preferences can be rationalized in the strong sense of counting as rational. For time preferences, this is done by adopting an interpretation akin to Camerer's in the opening quote of this chapter: self-control is an inherent constraint to decision making and optimizing under this constraint is rational. For risk preference, this logic is carried somewhat automatically by the structure of their dual model where the account of risk preferences is embedded into their earlier account of time preferences. Fudenberg et al. (2014) further show that their account of risk preference does not violate first-order stochastic dominance, which we saw in the previous chapter is *the* necessary condition of rationality in standard economics. Notice that their dual model therefore provide new arguments,

[17] Fudenberg and Levine (2011) are the only dual modelers deriving quantitative prediction from their dual model from calibration with one set of data to predict another set of data. But this is not really a marker of the standard versus behavioral economics contrast. Also, all the dual trend in *P*sychology was driven (post-2000) by the need to account for inter-individual differences. All dual modelers claim that their model can do that as well ('in principle'), but the only quantitative exercise by Fudenberg and Levine (2011, fn5 p.36) explicitly claim that representative agent predictions are better suited for economics.

namely from the internal constraints of self-control, for the rationality of violating the independence axiom that are different from Allais' initial arguments (i.e., the value of certainty for the avoidance of ruin, cf. the previous chapter). In Loewenstein and O'Donoghue's applications, it would be more appropriate to say that most of behavioral economics' challenges to risk and time preferences are *quasi*-rationalized, following Thaler and Shefrin's terminology. As they put it in an earlier version of their dual model: "[o]ur general modeling strategy is to assume that the deliberative system conforms to prescriptive models of decision making, such as expected utility, whereas the affective system incorporates many descriptive features of decision making identified in psychology" (Loewenstein and O'Donoghue 2007, p.4). In the published version (as well as in the previous ones) arguments for this strategy are further provided by references to the biological sources of normativity from neurosciences and animal studies discussed in the previous section (aside from casual empiricism and introspection). Therefore, their view seems to be that what counts as rational are the deliberative system's optima, that the individual would like to attain but cannot, though without going all the way down to affective system's optima. The individual is in-between, namely quasi-rational. For social preferences, the issue is more ambiguous with both dual models. Dreber et al. (2016) do not even implicitly mention the issue of normativity or rationality in their applications to social preferences. We can however speculate that the work of Fudenberg and Levine (2012a) on social preferences discussed in the previous chapter, on the need to have model of social preferences extendable under risk without respecting the independence axiom to capture some behaviors that are arguably rational, could be done through their dual model. But they have not work on the interaction between risk and other people yet. And as already mentioned, Loewenstein et al. (2015) suggest that the affective system's instability towards other people is not rational, but they carefully suggest that such instability regarding the motives at play in dictator games, which they do not treat, clearly has a deliberative dimension as well.

Summing up this subsection, we can note two main contrasts between Fudenberg and Levine's and Loewenstein and O'Donoghue's applications. First, besides the general primacy of time preferences in the sense of the previous subsection, there is a clear primacy of risk over social preferences in the following senses. For Fudenberg and Levine, this primacy is exhibited in the formal structure of the settings of their applications, i.e., the social preferences applications are

made within the structure of the bank/nightclub originally designed for risk preferences, itself an extension of the consumption-saving setting of their first application. This somewhat complex structure allows Fudenberg and Levine to account for interactions across dimension somewhat naturally, as their early bold predictions made in their 2006 paper shows. For Loewenstein and O'Donoghue, the primacy of risk over social preferences is exhibited by the differential number of pages and phenomena accounted for, along with their own introduction of their applications to risk (but not social) preferences as "natural" (Loewenstein et al. 2015, p.65). Finally, in terms of the positive/normative issue within dual models, both point to two different views of rationality as either a humanly attainable characteristic of decision making (Fudenberg and Levine) or as an ideal norm mostly beyond human reach (Loewenstein and O'Donoghue).

Conclusion

This subsection has compared how dual models achieved theoretical unification in their applications. We have seen that even the two couples of dual modelers (Bernheim and Rangel, Benhabib and Bisin) that did not focus on the behavioral phenomena discussed in the previous chapter played a non-trivial role in the current behavioral *versus* standard economics debates. The remaining three couples of dual modelers (Brocas and Carillo, Fudenberg and Levine, Loewenstein and O'Donoghue) achieve theoretical unification for time preferences in roughly the same way, namely by modeling the interactions between their two sub-individual entities through self-control so as to turn the original constant discount rate into a variable. Only two couples of dual modelers (Fudenberg and Levine, Loewenstein and O'Donoghue) tackle theoretical unification for risk and social preferences. But only Fudenberg and Levine have a theoretical structure that account quite straightforwardly for interactions across dimensions, albeit not for all the interactions discussed in the previous chapter (e.g., not between risk and social preferences). We have also further illustrated the tension between two possible interpretations of the normative/positive issue within dual models. Behavioral economists (Loewenstein and O'Donoghue) seem to follow the psychologists discussed in the previous section in taking rationality as an ideal norm that we are unlikely to attain empirically. And standard economists seem to view rationality as both an empirical characteristic of decision making and an ideal norm, with sources of normativity that seems more problematic (ignorance for Brocas and Carillo) than others

(willpower capacity for Fudenberg and Levine). Finally, it can be argued that the emerging implicit primacy of time preferences emerging in decision theory that was illustrated in the previous chapter is here complete and explicit.

Conclusion and transition: framing as the limit to theoretical unification

The main goal of this chapter was to conduct a comparative study of the five dual models altogether. It was done by focusing on how the import of a new language from *P*sychology, namely, the language of self-control, was informative to understand both the interdisciplinarity and positive/normative issues within dual models. These two issues raised themselves quite differently within each dual model. One difference was the extent to which dual modelers sought to contribute directly to *P*sychology. The two extrema of minimal and maximal contributions to *P*sychology are respectively represented by Fudenberg and Levine's and Loewenstein and O'Donoghue's dual models. Another difference concerned the scope of theoretical unification. Here, both Bernheim and Rangel's and Benhabib and Bisin's dual models represent the minimal extremum (addiction and consumption-savings), while Loewenstein and O'Donoghue's and especially Fudenberg and Levine's dual models represent the maximal extremum (time, risk and social preferences). A further difference concerned the positive/normative issue, with the following tension that was illustrated throughout (besides a plurality of sources of normativity). On the one hand, behavioral economists Loewenstein and O'Donoghue tend to side with psychologists' view that models of decision from standard economics provide a sound but unattainable set of norms of rationality. Their dual model is however not taken to be such a model, it is rather an in-between quasi-rational model of decision making. By contrast, the other (at least 'former') standard economists tend to stick to the view that models of decision in economics should picture rationality as an empirically attainable and indeed (most of the time) empirically attained set of norms of behaviors.[18]

[18] As already noted in one footnote of this chapter's introduction, this tension could be further analyzed through the philosophical literature on normativity and rationality in the cognitive (see Stein 1996) and social (see Turner 2010) sciences. Somewhat ironically, the conclusion of this chapter would be that the tension between the two ways of interpreting dual models with respect to rationality correspond to Stein's (1996) "pragmatic picture of rationality" and "standard picture of rationality", but it is the standard economists who hold the pragmatic one and the behavioral economists who hold the standard one.

Can we therefore conclude that one dual model is likely to occupy a central place in microeconomics (i.e., Camerer's suggestion in the introduction)? Arguably, this is not yet the case and even if it does happen, which dual models will hold such central place will depend on what microeconomists are most interested in. For instance, the core of Bernheim and Rangel's dual model is fit for the interests of those who practice normative economics; Benhabib and Bisin's for those interested more in microfoundations of representative agent models in macroeconomics; Brocas and Carillo's for those taking information as the central concept in microeconomics; Loewenstein and O'Donoghue's for those interested in furthering interdisciplinary communication with psychologists. Fudenberg and Levine's dual model is fit for the criterion set up in this chapter's introduction, which was in terms of unificatory power, especially with respect to the behavioral phenomena discussed in the previous chapter about the interactions within and crucially *across* the three dimensions of economic behaviors (over time, under uncertainty and regarding other people). Fudenberg and Levine's dual model is the only one to account for interactions across dimensions. Their theoretical unification is achieved through an explicit primacy of time over risk over social preferences. And new normative arguments from self-control as a mental constraint are derived to rationalize behavioral regularities that were previously considered as irrational. Another way of saying this is that part of the notion of consistency would change if their dual model were to hold a central place in microeconomics. The primacy of time preferences underlying such change is however in line with the latent normative primacy of time preferences in the previous chapter (cf. the ultimate justification of the independence axiom from dynamic consistency).

There is at least two ways in which Fudenberg and Levine's theoretical unification is not complete. First, all the types of interactions across dimensions discussed in the previous chapter are not explicitly accounted for yet (e.g., between risk and other people). But it would not be surprising if that happen in a future extension of their dual model. Second, there are a set of behavioral phenomena that are usually labeled 'framing' which Fudenberg and Levine repeatedly designate as *the* class of phenomena that is hard to capture formally. For instance,

> "Cues are obviously the key to understanding the effect of context in general, and framing in particular. The dual-self theory implies that it is the attention span of the short-run self that is relevant for determining what constitutes a "situation," which is the most difficult modeling issue in confronting these types of issues." (Fudenberg

and Levine 2006, p.1472)

Even more explicitly, Fudenberg claimed that "[u]nfortunately, framing and context are very difficult to capture in formal models, and are ignored in most of the more formal papers" (2006, p.699), and that "frames, cues, and mental accounts, for the time being they are a crucial but unexplained part of many behavioral analyses" (2006, p.700; see also p.708). These types of remarks are present in both Fudenberg and Levine's and other dual modeler's writings.[19]

That framing phenomena are hard to capture formally, even with the flexibility offered by dual models, might sound somewhat paradoxical because 'framing' was explicitly a key modeling ingredient of Shefrin and Thaler's (1988) original dual model. Shefrin and Thaler introduced framing as an instances of mental accounting (see 1988, p.610 and p.615), whereby different presentations of the same amount of money generate different marginal propensity to consume. Their favorite examples of framing are bonuses:

> "In a standard economic model, a completely anticipated bonus is simply income with another name. Thus the distribution of earnings into income and bonus would be considered irrelevant." (p.633)

In Shefrin and Thaler's (1988) paper, as well as in the references mentioned in the previous paragraph and footnote, different 'frames' or 'framing' of something (consequences, problems, amount of money etc.) are often said to be different *descriptions* of that something. But characterizing framing in this way is deeply ambiguous regarding whose descriptions we are talking about. It can be economists' descriptions, e.g., (1) "anticipated bonus is simply income with another name". But it could also be economic agents' descriptions, in which case there is a further complication that can be illustrated by the following contrast: (2) 'here is your promised €1000 bonus' – say the decision modeler to the decision maker – *versus* (3) 'I finally got that €1000 bonus' – say the decision maker to himself while doing his mental accounting. Whether it can be said that (1), (2) and (3) are equivalent in the sense of just being redescriptions of the same phenomenon is the kind of question that will be asked in the next chapter. By contrast with the present chapter, where we abstracted from the uses of language by economic agents to focus on the economists', in the next one we shall abstract the analysis from the economists'

[19] See Benhabib and Bisin (2008, p.325; see also Benhabib et al. 2010, p.206), Bernheim and Rangel (2007a, esp. pp.10-11, 35, 38, fn25 p.61, p.64, 66; see also Saez 2007 comments on on Bernheim and Rangel 2007a), Brocas (2012, p.307), Loewenstein and O'Donoghue (2004, p.39).

uses of language to focus on the economic agents', which we will argue is crucial for a full understanding of framing phenomena.

Chapter 4

Language as behavior: the structure of framing phenomena

> "Judgements about the intended meaning of utterances are themselves judgements under uncertainty" (Hilton 1995, p.266)

The previous chapter ended by pointing 'framing phenomena' as the limit to dual models' theoretical unification and reconciliation between behavioral and standard economics. This chapter scrutinizes framing phenomena to understand their empirical structure and theoretical implications for behavioral and standard economics. As can be already glanced in Fudenberg (and Levine)'s quotes ending the previous chapter, the issues underlying framing phenomena go well beyond dual models. We shall see that they extend to all formal modeling of individual decisions in economics (and maybe even beyond that), irrespective of the 'behavioral' or 'standard' style of modeling. The problem in trying to understand exactly how this is the case is that the uses of the words 'frame', 'framing' and 'framing effects' is getting broader and broader. That is, these are used to refer to an increasingly heterogeneous set of phenomena. The only element these phenomena have in common is that they consist in preference reversals triggered by different presentations of the same decision problem, i.e., 'the same' from the perspective of standard models of individual behaviors. Heterogeneity in framing phenomena comes from the various kinds of differences in the presentations. There can be different orders of presentation

of the objects of choice (first presenting the vegetables and then the hamburger or *vice-versa*), different ways of eliciting preferences (by choice, preference statement, cash equivalent, certainty equivalent etc.), different objects of choice that can be set as a "default" (chosen when no choice is actively made), etc. That heterogeneity shows that there are different ways of establishing equivalences between presentations of decision problems from the standard model's perspective.[1]

This chapter will not deal with all types of framing equally. It will focus on what is arguably one of the most paradigmatic, pervasive and formally challenging kinds of framing, namely the 'framing of consequences'. The classical example is known as the 'Asian Disease' from the original paper that introduced framing phenomena in decision theory back in 1981 by Tversky and Kahneman. We can take the occasion to illustrate the Asian Disease to introduce the main terminology and way decision problems will be presented throughout this chapter to dissect their empirical structure, notably using brackets:

> [*decision scenario*] Imagine that the U.S. is preparing for the outbreak of an unusual Asian Disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
> [*object of choice*] If Program A is adopted, 200 people will be saved.
> [*object of choice*] If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.
> [*decision question*] Which of the two programs would you favor?
> [*object of choice*] If Program C is adopted, 400 people will die.
> [*object of choice*] If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.
> [*decision question*] Which of the two programs would you favor?

Beside thinking about your preferences among programs A, B, C and D, try also to think about whether all the programs are *equivalent*, how you would construct other programs that you would consider to be equivalent, and whether you would present all these programs equivalently to the agent who is supposed to make the choice. You may wonder, 'But equivalent with respect to what?'. In a way, this is the central question of this chapter. Traditionally, experimental subjects are presented with *identical* decision scenario ("Imagine that...") and decision question ("Which of...") but only one of two *frames* that are supposed to be *equivalent*: either the 'save frame' where the choice set is $\{A, B\}$ or the 'die frame' where the choice set is $\{C, D\}$. Notice

[1] Textual evidence for the heterogeneity in the uses of 'frame', 'framing' and 'framing effect' can be found by reading, e.g., Bernheim and Rangel (2007a), Saez (2007), Salant and Rubinstein (2008), DellaVigna (2009), Holt and Laury (2014), among many others.

that $\{A, B\}$ and $\{C, D\}$ are identical in terms of their consequences because the numbers of people that die or that will be saved in the sure programs (A and B) and in the probabilistic ones (C and D) are numerically *equal*. The only differences lie in the descriptions of these consequences, i.e., in the *framing* of the objects of choice. A *framing effect* is usually observed here because subjects tend to choose A over B but D over C, i.e., they exhibit a preference reversal. In short, different descriptions of the same objects of choice reveal preference reversals. I want to emphasize that the framing of consequences discussed in this chapter involve different descriptions *in ordinary language*, not, e.g., in terms of graphical illustrations.

My use of the notions of equivalence, identity and equality is borrowed from philosopher Craig Dilworth (1988). According to Dilworth, "the view that identity is a relation may derive from its being confused with equality and equivalence, and that, in the case of the identity of a thing with itself, identity in effect marks the absence of relation" (1988, p.83). Roughly, on Dilworth's account, identity characterizes individuation of things, concepts, terms, or more generally of entities; equivalence characterizes relations between individuated entities; and equality characterizes identity of numerical values, either in its most abstract form in mathematics (especially in number theory), or as the result of measurement procedures in the empirical sciences. This distinction is useful to understand the structure of framing phenomena because the notion of equivalence is central yet often confused with identity in comments on the empirical structure of framing experiments. Furthermore, framing experiments involve numerical quantities, and the equality or inequality between some of these quantities is central in the theoretical explanations from economics or *P*sychology.[2]

[2]Dilworth (1988) defends a position that is controversial regarding the foundations of logic, but that does not really matter here because his distinction and vocabulary will be used for its virtues of clarification regarding the structure of framing experiments, *not* regarding the role of logic in the theories of psychologists and economists. Still some more details on his position are in order here. Individuation by identity is, on his account, marked by the absence of relations, and not by the presence of a reflexive relation, as is more traditionally conceived. Equivalence, by contrast, characterizes the very presence of (conceptual, linguistic, physical etc.) relations between distinct (conceptual, linguistic, physical etc.) entities, i.e., entities that have distinct identities. Furthermore, he argues that "[o]ne can speak of many sorts of things as being equivalent, and the notion is often used in the context of attaining some end", for instance: "certain actions may be equivalent – running twelve minutes is equivalent to walking one hour, when it comes to the burning of calories" (Dilworth, 1988, p.86). Finally, in $1 + 1 = 2$, the two 1s in the left side of the equations are identical and equal, and the mathematical expression $1 + 1$ is not identical to 2, though it is equal to it because both have the same numerical values, namely 2 (ibid, p.88). And "in the case of physical laws expressed by equations what are equated are measurable properties (parameters)" (ibid). In $U = IR$, which expresses Ohm's law whereby "in a closed circuit voltage is equal to current times resistance", "voltage is not being said to be identical with current times resistance – what is identical in this case, as in the purely mathematical case, is the numerical value of that which is represented on each side of the equation" (ibid).

Like Fudenberg and Levine along with the other dual modelers at the end of the previous chapter, plenty of standard and behavioral economists have commented on framing phenomena 'in passing', so to speak. From all these comments, the following big picture emerges. First, framing phenomena have not been subjected to any systematic study in behavioral or standard economics: ""framing" has yet to receive any complete treatment, though various insights have emerged over the years" (Quiggin 2014, p.713). Second, the framing effects in the framing of consequences are taken to violate "an invisible axiom of preference theory [...] that the way a choice is described should not influence its attractiveness ("*description-invariance*")" (Camerer 2004, p.386, my emphasis). Third, because of that, they supposedly don't fit within a utility maximizing framework; i.e., they are supposedly "mathematically intractable" (Tversky and Kahneman 1986, p.89). Fourth, framings of consequences are supposedly explained by nonstandard models of individual behaviors, especially prospect or cumulative prospect theory (see Tversky and Kahneman 1981; 1992). Fifth, violations of description invariance are taken as instances of irrationality; they are supposedly "normatively distasteful" (Tversky and Kahneman 1986, p.89).[3]

Beside these comments, there are also some contributions on the framing of consequences in economics, though they are few in numbers and they will be discussed in the next chapter as the related literature to the formal account presented there. In both these comments and contributions, economists tend to focus their discussion of the framing of consequences on Kahneman and Tversky's 1980s work. Yet other psychologists have, over the last twenty-five years or so, progressively refined the conditions under which this framing effect holds. The number of publication on this topic is fairly high, hence it is no surprise that in the few instances where economists have referred to them, some crucial features of the underlying framing effect are still missing. Accordingly, the main goal of this chapter is to contribute to the first point of the big picture from the previous paragraph. To do so, a systematic study of how the framings of consequences have been handled in *P*sychology will be conducted by taking the Asian

[3]Textual evidence for these claims can be found in the following references (pages indicate especially clear statements of one of the five points): Arrow (1982, pp.6-8), Smith (1985), Machina (1987, pp.144-146), Kreps (1988, p.197), Sugden (1991, p.759), Quiggin (1993, §14.5), Camerer (1995, p.652; 1999, p.10577), Rabin (1996, pp.46-47; 1998, pp.37-38; 2002, p.662), Starmer (2004 [2000], p.111-2, p.129), Luce (2000, p.8, pp.34-35, pp.216-217), Camerer and Loewenstein (2004, p.12, p.14), Samuelson (2005, pp.93-95), Varian (2005, pp.549-550), Thaler and Sunstein (2008, p.37), Kösegi and Rabin (2008a, p.1824), Loewenstein and Haisley (2008, pp.219), Schotter (2008, p.73), Wakker (2010, p.234, pp.241-2, p.250, p.265, p.350, pp.377-378), Bardsley et al. (2010, pp.130-1), Kahneman (2011, chap.34), Angner and Loewenstein (2012, p.663, p.668), Spiegler (2013).

Disease as a guiding benchmark for comparison with other framing phenomena. The issue of interdisciplinarity will be dealt with by conducting this systematic study from the perspective of economics regarding the second, third and fourth points. That is (and respectively), the demands of the implicit axiom of description invariance will be characterized, how it may be weakened to account for its violations formally (which is done in the next co-authored chapter) will be informally discussed, and how prospect theory cannot straightforwardly account for existing variations on the Asian Disease will be pointed out. Also, further experiments to better understand the framing of consequences through the Asian Disease will be suggested along the way. The positive/normative issue will be dealt with by scrutinizing how the entanglement of facts, values and conventions in the framing of consequences suggests some conditions under which the underlying framing effects is rational, hence arguing against a generalization of the fifth point.

The main argument of this chapter is that a full understanding of the framing of consequences requires a systematic account of the potential interactions between (1) how the decision modeler presents the decision problem and (2) how the decision maker represents this decision problem to himself. Indeed, Kahneman retrospectively acknowledged that an "important distinction between what decision makers do and what is done to them" was "blurred" by "[t]he use of a single term", i.e., "the label "frame"" in their pioneering contributions (2000, p.xvi). To mark this distinction throughout the chapter we shall call (1) 'external frames' and (2) 'internal frames'. The uses of language by economic agents is a condition of possibility for the interactions between external and internal frames to take place. Such uses of language is here considered as economic behaviors (hence part of the title of this chapter: *language as behavior*), as already argued in the first chapter of this dissertation. Indeed, the account of the interactions between external and internal frames given here is nothing but a development to what was called the communicative structure of choices there. To facilitate this development, we shall abstract from the uses of language by economists (in the sense used in the previous chapter, i.e., in his activity of theory making) to focus on the uses of language by (economic) agents in their activity of decision making. It will also be argued that, from the perspective of economics, it is more the structure of the objects of choice that needs to be modified than the structure of preferences, which should be able to be represented over *different* possible description*s* of *one*

same consequence. Hence, unlike existing discussions of framing effects, more attention will be devoted to the empirical structure of external frames than to the psychological interpretations of their impacts on internal frames.

This chapter is organized in three sections. The systematic study is first conducted through Kahneman, Tversky and other behavioral economists' works on framing (4.1), before moving to other psychologists who adopt a perspective that is quite compatible with the latter's (4.2), and finally to psychologists who share a perspective on the communicative structure of choices that is well in line with this dissertation's (4.3). The conclusion of this chapter draws some implications for other framing phenomena and regarding the notorious fuzziness and permissiveness underlying the notion of consequence used in economics.

4.1 Pervasiveness and empirical subtleties of framing phenomena

The framework that will be used to conduct the systematic study of framing phenomena is constituted by three distinctions. Two of them have already be introduced. The first one is between identity, equivalence and equality, and constitutes the main originality of the framework in so far as it has never been used (to the best of my knowledge) to study framing phenomena. The second one is between external and internal frames. The last one comes from psychologist Deborah Frisch (1993, p.399), who coined the distinction between "strict" and "loose" framings (her work is discussed in the third section). Strict framings are about "pairs of problems that involve a redescription of the exact same situation" (ibid) as in the Asian Disease. Loose framings are about "pairs of problems that aren't exactly the same, but which are equivalent from the perspective of economic theory" (ibid) as in much of the framing effects discussed in this subsection. One goal of the systematic study of framing phenomena is to highlight the various conventions from economic theory that allows the establishment of equivalence between pairs of decision problems. Throughout, I will mainly focus on decision modelers' external frames in framing experiments, only mentioning the decision makers' modal preferences (using bold characters as in chapter 2). As we shall see, framing phenomena are pervasive in the sense that they occur under uncertainty (4.1.1), under certainty (4.1.2), regarding other people (4.1.3) and over time (4.1.4). Theoretical accounts of the internal frames (4.1.5) and implications for the

positive/normative issue (4.1.6) are briefly discussed for all the framing phenomena altogether at the end.[4]

4.1.1 Under uncertainty

We can first contrast the Asian Disease with the following "isolation effect" (Kahneman and Tversky, 1979, p.273):

> [Decision scenario] In addition with whatever you own, you have been given €1000.
> [decision question] You are now asked to choose between
> A: 50% chance of winning €1000 and 50% chance of winning nothing
> **B**: 100% chance of winning €500 for sure
>
> [Non-identical decision scenario] In addition with whatever you own, you have been given €2000. [identical decision question] You are now asked to choose between
> **C**: 50% chance of losing €1000 and 50% chance of losing nothing
> D: 100% chance of losing €500 for sure

This is a loose framing effect because the first 'gain frame' and the second 'loss frame' are not redescriptions of an identical situation. Indeed, the decision scenario needs to change across frames for the choice sets $\{A, B\}$ and $\{C, D\}$ to have equivalent consequences from the perspective of economic theory. The equivalence is established through the convention of counting as consequences, not the mere outcome of a choice, but the consequence of such outcome on the decision maker's wealth. Hence the necessity to integrate the gains or losses with the endowment given in the decision scenario. Once this integration is done, it yields numerically equal quantities, making the consequences equivalent from the perspective of economic theory. The consequences of choosing A or C, and B or D are, by themselves, not identical as they involve either gains or losses; only the consequence of wining nothing or losing nothing is identical in the two frames. By contrast, in the strict framing of the Asian Disease, the decision scenario is identical across the 'save frame' and the 'die frame'. It does not need to change for the consequences to be equivalent from the perspective of economic theory.

In their 1981 paper Tversky and Kahneman provided a tripartition of external frames, into the framings of "acts", of "contingencies", and of "outcomes", the latter is taken as synonymous with 'consequences' in this chapter. Both the isolation effect and the Asian Disease are respectively loose and strict framings of consequences. To illustrate the contrast, consider the

[4] Parts of the developments in this section have been used in Jullien (20016a).

following loose framing of acts (Tversky and Kahneman, 1981, p.454), where the objects of choice A, B, C, D in the first frame are combined through the logical operator & to make new objects of choice $A\&D, B\&C$ in the second frame:

> [decision scenario & decision question] Imagine that you face the following pair of concurrent decisions. First examine both decisions, then indicate the options you prefer.
> Decision (i). Choose between:
> **A.** a sure gain of €240
> B. 25% chance to gain €1000, and 75% chance to gain nothing
> Decision (ii). Choose between:
> C. a sure loss of €750
> **D.** 75% chance to lose €1000, and 25% chance to lose nothing
>
> [non-identical (because absence of) decision scenario & identical decision question] Choose between:
> A & D. 25% chance to win €240, and 75.% chance to lose €760.
> **B & C.** 25% chance to win €250, and 75% chance to lose €750.

At first sight, there is no straightforward relation of equivalence between the two frames, but a relation of inclusion. In the first frame, two choices have to be made, one in $\{A, B\}$ and one in $\{C, D\}$, so that the choice set is in fact equivalent to $\{A\&C, A\&D, B\&C, B\&D\}$, in which the second frame $\{A\&D, B\&C\}$ is included. From the conventions of decision theory, however, one can see a loose equivalence between these two frames by considering that the part of the former not included in the latter, i.e., $A\&C$ and $B\&D$, are irrelevant. The former because it is a strict loss. And the latter because you would have a non-binary lottery with three consequences, i.e., +€1000 (with probability $\frac{9}{16}$), €0 (with probability $\frac{6}{16}$) and -€1000 (with probability $\frac{1}{16}$). Arguably, the irrelevance of the latter is more disputable than the irrelevance of the former. More generally, the point is that if one considers that $A\&C$ and $B\&D$ are relevant, then he can argues against Tversky and Kahneman that we are in the presence of a framing effect here.

In the following framing of contingencies (Tversky and Kahneman, 1986, pp.S263-4), notice how the blue marbles in the first ($\{A, B\}$) frame disappear but not their consequences, as they are redistributed into the yellow marbles in C and into the green marbles in D, and how the green and red marbles in B have the same consequences and have been combined into the red marbles in D:

[decision scenario] Consider the following lotteries [...]. [decision question] Which lottery do you prefer?

Marbles	*White*	*Red*	*Green*	*Blue*	*Yellow*
Option A	90%, €0	6%, win €45	1%, win €30	1%, lose €15	2%, lose €15
Option B	90%, €0	6%, win €45	1%, win €45	1%, lose €10	2%, lose €15

[identical decision scenario & decision question]

Marbles	*White*	*Red*	*Green*	*Yellow*
Option C	90%, €0	6%, win €45	1%, win €30	3%, lose €15
Option D	90%, €0	7%, win €45	1%, lose €10	2%, lose €15

Here, the equivalence is achieved by combining marbles that yield the same distribution of expected values. The blue and yellow marbles in A yield the same expected consequences, and are combined into the yellow marbles in C. The red and green marbles in B yield the same expected consequences, and are combined into the red marbles in D. Thus the overall expected values and distributions of consequences of both lotteries are preserved between frames, which make them equivalent from expected utility theory.

These framings of acts and of contingencies are loose ones because the equivalence from the point of view of expected utility theory across frames does nonetheless underlie consequences that not physically identical across frames. That is, for each framings, the two frames refer to non-identical situations. In the framing of acts, the opportunity to make one binary choice *disappears* (along with the decision scenario) and sums of money are *added*. In the framing of contingencies, marbles are *removed* and *added*. Hence, changes in the written words (that are thus non-identical across frames) imply changes in the physically implemented experimental setups (that would thus be non-identical across frames). In the Asian Disease, by contrast, there are physical identity of the consequences: 200 people live and 400 people die in both A and C, and this reasoning carries over to the probabilistic consequences. The non-identical features in the Asian Disease's objects of choice do not concern the consequences but only the words describing them. To illustrate the contrast more vividly, if you remove the words "blue: lose €15", you ought to remove the blue marbles in the underlying experiment, but the redescription of "200 people will be saved" (out of 600 people) into "400 people will die" (out of 600 people) only implies changes in the written words used in the experiments.

[decision scenario] You receive £50

[decision question] Choose between the gamble or the **sure option**

[identical decision scenario] You receive £50

[identical decision question] Choose between the **gamble** or the sure option

Figure 4.1: Stricter framing effect from De Martino et al. (2006)

Recall from chapter 2 that, intuitively, one lottery first-order stochastically dominates another one if you can win more money without taking more risk. Notice that the external frames in these framings of acts and of contingencies involve such relations of first-order stochastic dominance: $B\&C$ dominates $A\&B$ for the acts, and B and D dominate respectively$A$ and C for the contingencies. First-order stochastic dominance is the most important type of *relation between lotteries* ('in the world', i.e., in external frames) about which expected utility theory has something to say concerning individuals' preferences ('in the head', i.e., in internal frames) regardless of risk attitudes: namely, individuals should prefer the dominating lottery over the dominated one. Unlike the frames being either *positive* ('save') or *negative* ('die') in the Asian Disease, they are said to be '*transparent*' ($\{A\&D, B\&C\}$ for the acts, $\{A, B\}$ for the contingencies) when they make the relation of dominance clear and '*nontransparent*' ($\{A, B, C, D\}$ for the acts, $\{C, D\}$ for the contingencies) when it is not clear. Indeed, in the transparent frames, decision makers do not violate preference for first-order stochastic dominance. There was no violation of preference for first-order stochastic dominance in the Asian Disease for the simple reason that there is no such relations of dominance within its choice set.

The last framing effect under risk to be contrasted to with the Asian Disease comes from by neuroeconomists Benedetto De Martino, Dharshan Kumaran, Ben Seymou and Raymond Dolan (2006, p.685); it is reproduced in figure 4.1.

With respect to the Asian Disease, there is a further invariant dimension here: the pictorial design allows to use identical words in the presentation of the probabilistic object of choice

across frames (which in both cases read '1/3 probability of keeping the £50 and 2/3 probability of losing the £50'). In this sense, this is a stricter framing of consequence than the Asian Disease under risk, and suggests that there seem to be a specific role played by the sure consequence in it (the next two sections will provide strong support to this point).[5]

4.1.2 Under certainty

Framing phenomena under certainty have indeed been observed since the beginning of Tversky and Kahneman's work on the topic. Here is a well-known example of a loose framing of consequences under certainty (Tversky and Kahneman, 1981, p.457):

> [decision scenario:] Imagine that you have decided to see a play where admission is €10 per ticket. As you enter the theater you discover that you have lost a €10 bill.
> [decision question:] Would you still pay €10 for a ticket for the play? **Yes** No
>
> [non-identical decision scenario:] Imagine that you have decided to see a play and paid the admission price of €10 per ticket. As you enter the theater you discover that you have lost the ticket. [...]
> [non-identical decision question:] Would you pay €10 for another ticket? Yes **No**

The framing effect here is obviously a loose one: by contrast with the Asian Disease, the frames are not redescriptions of an identical situation (one is a situation where a bill is lost, the other where a ticket is lost). The equivalence is established through the convention from economic theory that sunk costs (whatever the form they take) should not affect economic choices. In both frames there is an equal sunk cost of €10, which does not affect choices in the first 'lost bill frame', but it does so in the second 'lost ticket frame'.

There are acknowledged mutual influences between Kahneman and Tversky's work on framing (especially under certainty) and Thaler's work on mental accounting (see, e.g., Kahneman and Tversky 1984, p.347; Thaler 2008, p.12). Thaler's work on mental accounting indeed involves the construction of choice experiments triggering framing phenomena. For instance, in the following one, he presents subjects with pairs of decision scenarios, in each of which "two

[5] As far as I know, Hollard et al. (2016) provide the only experimental investigation of this stricter framing effect by economists. Other phenomena usually labeled as 'framing' under risk within the province of standard and behavioral economics include notably some in the literature on mental accounting discussed below (see Thaler and Johnson 1990, sects.4-5), others related to the writing of instruction and presentation of lotteries in measurement of risk preferences (see Holt and Laury 2014), and famous violations of procedure invariance (see Slovic and Lichtenstein 2006).

events occur in Mr. A's life and one event occurs in Mr. B's life", events that are explicitly "intended to be financially equivalent" (Thaler 1985, p.202):

> "[decision scenario]Mr. A's car was damaged in a parking lot. He had to spend €200 to repair the damage. The same day the car was damaged, he won €25 in the office football pool.
>
> [non-identical decision scenario]Mr. B's car was damaged in a parking lot. He had to spend €175 to repair the damage."
>
> [decision question]Who was more upset? A, **B**, no difference (ibid, p.204)

Besides the obvious looseness of this framing effect, by contrast with the Asian Disease (and) and all other experiments discussed so far, here it can be argued that Thaler takes whole frames to be the objects of choice. In a way, deciding who is the more upset is tantamount to ranking the two decision scenarios, which are therefore the objects of choice, the "no difference" being explicitly identified as "emotionally equivalent" (p.202).[6]

4.1.3 Regarding other people

The following framing effect is an early instance of a framing phenomena related to social preferences (Kahneman, Knetsch, and Thaler, 1986a, p.731):

> [decision scenario] A company is making a small profit. It is located in a community experiencing a recession with substantial unemployment but no inflation. There are many workers anxious to work at the company. The company decides to decrease wages and salaries 7% this year.
> [decision question] Acceptable **Unfair**
>
> [non-identical decision scenario] A company is making a small profit. It is located in a community experiencing a recession with substantial unemployment and inflation of 12%. There are many workers anxious to work at the company. The company decides to increase salaries only 5% this year.
> [identical decision question] **Acceptable** Unfair

There is a change of real income (i.e., purchasing power) that, in both frames, is equivalent to a 7% decrease. From the conventions of economic theory, economic behaviors should be based on

[6] We shall see the theoretical side of Thaler's mental accounting later, but note that in the original 1985 paper as well as in further extension, not all mental accounting phenomena are about decisions under certainty (see Thaler and Johnson 1995; Thaler 1999; Shafir and Thaler 2006; Thaler 2015, part III). Another topic where framing under certainty has been developed within the province of standard and behavioral economics is the behavioral study of money illusion (see Shafir, Diamond and Tversky 1997), which is also related to social preferences as the next decision problem also illustrate. An econometric study of the effects of advertisement in the field by Bertrand et al. (2010) on loan take out observed classic framing phenomena that can be thought of as being in-between under certainty and over time.

real, not nominal, prices. Hence social preferences for fairness should not be influenced by money illusion, as it is the case here. But here again, the framing effect is a loose one: by contrast with the Asian Disease the two frames describe two situations with non-identical consequences. Not only the amount of money written in the workers' bank account would differ, but so would the speech acts from the company, as declaring a decrease or an increase in wages do not have identical propositional contents. The reasons behind both speech acts are however identical, i.e., "a recession with substantial unemployment" and "There are many workers anxious to work at the company", as are the consumption bundles available in both frames.[7]

4.1.4 Over time

Finally, here is an instance of framing phenomena in intertemporal choices (Loewenstein 1988):

> [decision scenario] For *every* pair, please circle the choice you prefer
> (1) **€7 in 8 weeks** vs. €5 in 1 week
> (2) **€7 in 8 weeks** vs. €5.25 in 1 week
> (3) **€7 in 8weeks** vs. €6.25 in 1 week
> (4) €7 in 8 weeks vs. **€6.50 in 1 week**
>
> [identical decision scenario]
> (4′) **€7 in 1 week** vs €7.50 in 8 weeks
> (3′) **€7 in 1 week** vs €7.75 in 8 weeks
> (2′) **€7 in 1 week** vs. €8.75 in 8 weeks
> (1′) €7 in 1 week vs. **€9 in 8 weeks**

Some pairs are equivalent from the theoretical convention of constant discounting which implies that the minimum premium the decision maker is willing to be paid for incurring a delay on a given consequence should be numerically equal to maximum cost he is willing to pay for speeding up this consequence (see Loewenstein 1988, p.203). Under this theoretical convention,

[7] Notice that in the previous framing problems, the consequential equivalence arguably requires more economic assumptions, e.g., all loans (especially real estate's) are indexed on inflation and that all prices (not only the average price) grow with inflation (I thank Ismaël Rafaï for pointing this to me). Closer to how social preferences have been discussed previously in this dissertation, Dreber et al. (2013) tried to observe a framing phenomena in the dictator game. One manipulation involves loose framings: 100 cents are given either to the decision maker in a 'give frame' or to the other in a 'take frame' but the allocation is done in both cases by the former. Another manipulation involves strict framings: only the label of the game changes in the instruction, either the "Giving Game" or the "Keeping Game" (in some conditions the entire decision problem, including the description of the consequences in the objects of choice, remains identical. They do not observe any framing phenomena, by contrast with the effects of these manipulations on other games with a game theoretical structure. The original study introducing framing of acts in public good games is from economist Andreoni (1995). The one introducing the framing of label is from psychologists Liberman et al. (2004) who relabeled prisoner dilemma games either the "Wall Street Game" or the "Community Game". A recent contribution in economics observing framing effects in both these types of framing is from Dufwenberg et al. (2011).

(1) and (4) in the 'speed up frame' are respectively equivalent to (1′) and (4′) in the 'delay frame' without framing effects because the preferences in the former two respectively imply the preferences in the latter two. But (2) and (3) are respectively equivalent to (2′) and (3′) with a framing effect because the preferences in the former two respectively imply the reverse of the preferences in the latter two. Unlike in the Asian Disease, this is however not a strict framing effects since, obviously, the decision situations are not identical across both frames.[8]

4.1.5 Theoretical accounts of internal frames

We can now briefly turn to the internal frames as they features in the theoretical accounts of the authors who demonstrated the framing phenomena that have been presented. The underlying accounts of internal frames rely mostly on prospect theory, especially on the value function (cf. chapter 1).

The explanation is quite straightforward for the recession with/without inflation and lost ticket/bill problems, as the reference point does not change across decision problems. In the former, the reference point is one's current *nominal* wage. A decrease in it without inflation is seen as a loss (which it is in both nominal and real terms) and loss aversion steers social preferences towards unfairness, i.e., the decision would be rejected if given the opportunity. An increase in it with inflation is still seen as a gain (which it is not in real terms), which therefore seems like an acceptable, fair, decision. In the latter, the reference point is to buy a ticket to see a play. Buying it for the first time is seen as a gain in utility from the play to be seen which is superior to the disutility from the €10 costs, and the lost €10 bill is not considered as part of the same economic decision. Losing it and having to buy it for the second time is seen as a part of the same economic decision, which makes (especially through loss aversion) the disutility from the €20 costs in ticket superior to the gain in utility from the play to be seen. Thaler's work on mental accounting is a theoretical refinement, conceptually and formally,

[8]The framing phenomena presented here is reconstructed from Loewenstein's mean results and instructions in his appendix. Much of the framing phenomena in intertemporal choices is usually discussed about the strong violations of procedure invariance in the elicitation of time preferences (see Frederick et al. 2002; and Manzini et al. 2014 for a recent contribution). There are however a few experiments that are closer from the one discussed here. For instance, Benhabib et al. (2010) use different framings of willingness to accept. An noteworthy contribution in psychology are Kirby and Guatsello's (2001) experiments who implement Ainslie's notion of bundling (which he himself refer to as an instance of 'framing') central to his theoretical account of hyperbolic discounting.

of this latter kind of explanation from prospect theory. In these refinements, money or utility are not necessarily fungible and Thaler extends the value function ($u(.)$) to two consequences that can be received jointly, i.e., at the same time. Roughly, Thaler provide various conditions under which the joint receipts of two consequences will be *integrated*, i.e., $u(x, y) = u(x + y)$, or *segregated*, i.e., $u(x, y) = u(x) + u(y)$, with the relaxation of fungibility allowing to treat cases where both would not yield the same utility, i.e., $u(x + y) \neq u(x) + u(y)$. This depends, among other features, on whether the consequences are gains (i.e., x, y) or losses (i.e., $-x, -y$) or both, in which case whether the gains are greater than the losses or *vice-versa* matters. The car damage example is easily explained by firstly setting the reference point to zero for both Mr. A and Mr. B. And secondly by assuming A segregates his mixed consequences thus yielding less disutility than B's unique negative consequence, i.e., $u_A(-€200) + u_A(€25) > u_B(-€175)$. Hence the latter is more upset as loss aversion is greater for consequences closer to the reference point.

Loewenstein (1988) explains framing phenomena over time simply by using prospect theory's value function within an otherwise standard model of discounting. The reference point changes across both external and internal frames, from '€7 in 8 weeks' to '€7 in 1 week'. Hence, the disutility from the monetary losses involved in the first frame are higher than the utility from the monetary gains in the second frame despite both amounts of money being numerically equal in absolute value. Thus a shift in the reference point and loss aversion explain the differences in elicited discount rates without postulating different time preferences.

Contrasting the isolation effect with the Asian Disease is especially instructive of the degree of liberty offered by potential but not necessary changes in reference point within prospect theory. In the isolation effect, the reference point changes across external frames (from +€1000 to +€2000), but remains invariant in the internal frames. That is, the decision maker evaluates both frames with a same reference point of zero and hence sees the consequences in the 'gain frame' as involving gains (hence choosing the sure consequence), and the ones in 'loss frame' as involving losses (hence choosing the risky consequences). By contrast, in the Asian Disease, the reference point is invariant across external frames ("600 people are expected to be killed"), but changes in the internal frames. As Kahneman and Tversky put it, $\{A, B\}$ "implicitly adopts as a reference point a state of affairs in which the disease is allowed to take its toll of 600 lives",

while $\{C, D\}$ "assumes a reference state in which no one dies of the disease" (1984, p.341). But who "adopts" and who "assumes" here? This is an instance of the confusion between external and internal frames retrospectively acknowledged by Kahneman (2000). It is only because the Asian Disease is carefully designed as it is by the decision modeler (here Kahneman and Tversky the experimenters) that different reference points are induced. In this respect and by contrast with the isolation effect, the Asian Disease shows that the framing of the consequences within the choice set can, by itself, give specific meanings to the decision scenario interpreted as shifts of the reference point under prospect theory. De Martino et al.'s (2006) stricter framing of consequences, which is explained through prospect theory exactly as is the Asian Disease, shows *the key role of the framing of the sure consequence* for inducing shifts in the reference point. In the isolation effect, $\{A, B\}$ and $\{C, D\}$ are, taken in isolation, given gains and losses respectively; so it is not surprising that they are *seen as* gains and losses respectively. By contrast, in the Asian Disease's external frames, the sets of consequences taken in isolation are neither gains nor losses, they are mixed gains and losses. But it seems pretty obvious that $\{A, B\}$ are seen as gains and $\{C, D\}$ as losses in the internal frames, because of the semantics of "saved" and "die" respectively. This contrast highlights the difference between establishing an equivalence between frames through a theoretical convention, i.e., integration of gains an losses with the endowment, and through an atheoretic convention, i.e., the semantics of "saved" and "die" in ordinary language uses. Though both equivalences hold for the decision modeler, they don't for the decision maker.

The original (1979) weighting function also plays a role in the theoretical explanations of framing phenomena from the previous paragraph by exacerbating the risk seeking/risk aversion asymmetry across frames, its role is not as crucial as for both the framings of acts and (especially) of contingencies. In the framings of acts, the weighting function also exacerbate the asymmetry in risk preferences within the first frame but it does so up to a point leading to violation of first-order stochastic dominance (by choosing $A\&D$ over $B\&C$). In the framings of contingencies, the weighting function plays the crucial explanatory role through its property of subadditivty. The weights attributed to the probabilities of getting €45 by drawing a red marbles (with probability .06) *and* a green marble (with probability .01) in B of the transparent frame is superior to the weight attributed to the probability .07 of getting €45 by drawing a red marble in D of the

nontransparent frame, i.e., $\pi(.06) + \pi(.01) > \pi(.07)$, explaining the shift of a preference for B to a preference against D (Tversky and Kahneman, 1986, p.S263).

Finally, DeMartino et al.'s (2006) account of internal frames is based on the observation that the modal pattern of preferences exhibiting the framing effect is correlated with higher activation of the amygdala (than other patterns of preferences), a brain region usually associated with the experience of emotions. They accordingly emphasize the role of emotions in decision making and cast they results within a dual-system account of internal frames.

4.1.6 The positive/normative issue in framing phenomena

Despite the variety of framing phenomena sharing a fairly clear and common explanation from prospect theory, their status regarding the positive/normative issue are less clear and different. We shall be brief on the ones related to social and time preferences not only because there is not much to say but also because the rest of this chapter will focus on framings under certainty and uncertainty. The same standard position as the one highlighted in chapter 2 still holds for framings related to social preferences: 'fairness' judgments are studied only in so far as they are recognized as such by economic agents. Economists explicitly refuse to judge whether the content of such judgments are indeed fair or unfair, can be deemed rational or irrational (see esp. Kahneman et al. 1986a, p.729), though it is emphasized throughout that they violate standard models of individual behaviors. Likewise but more maybe surprisingly for time preferences, Loewenstein (1988) makes no comment on the normative implications of framing phenomena in intertemporal choices. Such implications have however been drawn from start in Kahneman and Tversky's work on framing and Thaler's on mental accounting (see the references mentioned above). This is notably the case in the following passage from Tversky and Kahneman (1981), which is also relevant for (and is indeed about) framings under certainty and uncertainty:[9]

> "Further complexities arise in the normative analysis because the framing of an action sometimes affects the actual experience of its outcomes [...]. [D]eliberate manipulation of framing is commonly used as an instrument of self-control [here a footnote references Strotz, Ainslie, Elster, and Thaler and Shefrin]. When framing

[9] See Prelec and Loewenstein (1998) along with Read et al. (1999) for further reflections of Loewenstein on framing in intertemporal choice connected with mental accounting.

> influences the experience of consequences, the adoption of a decision frame is an ethically significant act." (Tversky and Kahneman 1981, p.458)

This passage underlies two related tensions (both retrospectively acknowledged by Kahneman 2000), between the two types of frames and between two views about the notion of equivalence, which will be scrutinized in the rest of this chapter. The tension between the internal and external senses of a frame transpires in every sentences of the passage. Whose act of framing is ethically significant? Is it the decision modeler's choices of an external frame over another? In contemporary behavioral economics terms, this corresponds to the activity of 'choice architects' discussed in chapter 1, who for instance designs precommitment devices to solve other agent's self-control problems (e.g., pension plans, cf. chapter 2). Or is the ethically significant act the decision maker's choices of an internal frame over another? This corresponds to both the strategy of consistent planning (cf. chapter 2) or 'internal influences' (cf. chapter 3) to cope with self-control problems (e.g., an ethically significant internal frame to avoid succumbing to temptation consists in focusing as much as possible on the future bad consequences). From the arguments of chapter 1 about the communicative structure underlying most economic choices, it does not make much sense to consider both questions separately. The ethical significance, i.e., normative implications, of external and internal frames cannot be fully understood if not studied in interaction. Was the decision maker aware of his self-control problems before encountering the external frame? What was the intentions of the decision modeler in choosing one frame over another (compare: a friend recommending a product *versus* an ad for this product)? These issues underlie the interactions between internal and external frames. Kahneman and Tversky's theoretical focus was exclusively on internal frames. Besides explicit acknowledgment from Kahneman (2000), the change in label from the "editing phase" of the 1979 version of prospect theory to the "framing phase" of the 1992 cumulative prospect theory is a marker of this theoretical focus.

The other tension is, as Kahneman (2000, p.xv) puts it, between a "theory-bound" view and a "lenient" view of equivalence – Kahneman (2000) emphasizes that they (with Tversky) switched from the former to the latter in the 1980s. The theory-bound view is essentially the one displayed so far: equivalences are established from the theoretical conventions of economic theory. By contrast, in the lenient view, they are established by "the decision maker [...] after due

consideration of both problems" (ibid). Therefore and in principle, it is possible for a decision maker to rationally view two frames as non-equivalent and maintain his preference reversals even after the equivalence from economic theory has been explained to him. As emphasized by Tversky and Kahneman (1981, p.458), his justification can be hedonistic (he derives more utility from a given consequence under one frame than under the other) or ethical (one frame can fit some standards better than another frame). A tension arises in practice because such possibility, which is a key implication of the lenient view of equivalence, never arises in discussions of framing effects by Kahneman, Tversky and Thaler.

Arguably and from the perspective of this dissertation, this is sometimes quite sound, as for the framings under certainty, and of acts and contingencies under uncertainty discussed above. For instance, Kahneman (2011, p.371) includes the 'lost bill frame' in the framing under certainty in the "good frames" that lead to "more reasonable decisions" because "[h]istory is irrelevant", as the standard justification from the theoretical convention that sunk costs should not matter holds it. The discussion ends as follows:

> ""Would you have bought tickets if you had lost the equivalent amount of cash? If yes, go ahead and buy new ones." Broader frames and inclusive accounts generally lead to more rational decisions." (ibid)[10]

Within the constraints of the decision scenario of this experiment, there is indeed not much to be argued against that conclusion. Few would find the following justification convincing: 'I would be so upset to have lost the ticket that I will not enjoy the play, which does not happen if I had lost €10'. That there is no convincing justification comes mostly from the fact that there is no communicative structure in the decision. In a sense, within the real-world counterpart of the experiment you would be both the decision modeler and the decision maker for the choice of whether or not to buy a ticket, i.e., you construct your own counterfactual to make up your mind.

For the framings of acts and of contingencies leading to violations of first-order stochastic dominance, there is also not much to be argued for the rationality of preferring less to money to more money. Doing so would require a lot of contextual factors which are not involved in the lab experiments as witnessed by the 100% of the subjects who choose the dominant lottery under

[10]Thaler (2015, part III) holds a similar position on mental accounting phenomena.

transparent framing (see also MacCrimmon 1968, sect. V). Demonstrations of such framing effects were indeed central to Tversky and Kahneman's claims that "the dream of constructing a theory that is acceptable both descriptively and normatively appears unrealizable" (1986, p. S272). This is so because framing effects are pervasive (hence should be descriptively accounted for) but can lead to violations of first-order stochastic dominance (hence cannot be normatively defended as rational). One implication of such claims was to take the principle of description invariance as "normatively indispensable" (ibid) or "normatively unassailable" (Tversky and Koehler 1994, p.565).[11]

But it can be argued, especially from the systematic study conducted above, that description invariance does not have the same normative implications in all framing effects, mainly because *strict* framings cannot lead to violation of dominance. After all, a standard economics account (i.e., no value judgment on the content of preferences) could be given to strict framing effects, whereby they are not necessarily irrational as long as any arguments are made against the rationality of a preference for one description of a consequence (or of a decision problem) over another description of the same consequence (or of the same decision problem). Kahneman and Tversky have indirectly made such arguments. On the one hand, they mentioned some casual observation of decision makers who were explained the equivalence between frames in the Asian Disease and were then conflicted between keeping their preference reversals and respecting description invariance *at the same time* (e.g., Kahneman and Tversky 1984 p.343). On the other hand, Kahneman (2011, chap.34) discusses strict framings of consequences in the Asian Disease and De Martino et al's (2006) altogether by using the expression "emotional framing" over and over again, referring to what's going on in the Asian Disease as "empty intuition". He again reports the casual observation that "when people are confronted with their inconsistency" and are asked "["n]ow that you know these choices were inconsistent, how do you decide?" The answer is usually embarrassed silence" (Kahneman 2011, p.369). In short, strict framings of

[11] Notice that there is an apparent paradox here. Recall from chapter 2 that Kahneman and Tversky amended their 1979 prospect theory into their 1992 cumulative prospect theory in order to avoid violations of first order stochastic dominance. Recall also that the latter is the main theoretical convention in decision theory from which value judgments of rationality and irrationality are derived. Given that Kahneman and Tversky never claimed that cumulative prospect theory has a normative dimension, and that they had already demonstrated factual violations of first-order stochastic dominance prior to 1992 (the framings of contingencies and of acts), why did they wish to analytically prevent such violations? The paradox is however only apparent because these violations occur in these very specific situtations involving (nontransparent) framings, while the 1979 implications of such violations were systematic.

consequences are "illustration of how the emotion evoked by a word can "leak" into the final choice" (Kahneman 2011, p.366). The main problem with this position is that it virtually ignores more than two decades of research by other psychologists on the Asian Disease and other strict framings going beyond casual observation, i.e., doing experiments on the processes by which decision makers judge frames to be equivalent or not. By scrutinizing their work, we shall argue that something is indeed leaking in framing effects, and that is not just emotion but also information about the decision situation that can be relevant in some cases.

Conclusion

We can briefly conclude this section by noting the contrast between framing effects and the challenges to standard economics such as Allais' paradoxes seen in chapter 2. The latter's external frames were pairs of decision problems with unequal consequences across the two problems constitutive of a given pair. Despite this inequality, there was certain numerical relations between these pairs such that the modal preferences were deemed 'inconsistent' because of their violations of theoretical conventions such as the independence axiom. These modal preferences were not inconsistent anymore within theories that relaxed these theoretical conventions (e.g., rank-dependent utility theory), which was normatively justified by the work of Allais, MacCrimmon, Quiggin and others. By contrast, the framing effects discussed in this section are pairs of decision problems with equal consequences across the two problems constitutive of a given pair. In loose framing effects, these equalities are established through various theoretical conventions, each of which have more or less defensible economic logic. In strict framing effects, these equalities are established through atheoretic conventions from ordinary language uses, which have been argued *ex post* (to the demonstration of framing effects) to constitute an implicit axiom of the standard model, namely 'description invariance'. Description invariance is usually considered to be 'obviously' rational without justification by an economic logic beyond common sense. In short, the uses of ordinary language constitutive of the external frames in Allais' paradoxes are not as crucial to the underlying behavioral phenomenon as they are in strict framing effects.

The distinction between loose and strict framing effects, and especially the specificity of the latter within the distinction, are worth investigating because even psychologists specialized in

the matter have made confusions with non-trivial implications (see Fagley 1993).

4.2 Strict framings beyond Kahneman and Tversky

There is an impressive number of contributions on strict framing effects in *P*sychology, as the sheer number of reviews and meta-analyses testifies (i.e., Kühberger 1997; 1998; Kühberger et al. 1999; Levin et al. 1998; Piñon and Gambara 2005; Maule and Villejoubert 2007; Keren 2012; 2011a; Mandel and Vartanian 2012; Takemura 2014, Part V). As emphasized in the most cited of the latter, within the existing variation in framing effects, "the likelihood of obtaining choice reversals [is] directly related to the similarity between features of a given study and features of Tversky and Kahneman's (1981) original "Asian disease problem."" (Levin et al. 1998, p.157). The effect is robust to changes in scenario and type of consequences, e.g., money instead of lives. Hence, it makes sense to focus on the structure of the Asian Disease's external frame to guide potential theoretical developments. This section comments on a selective set of results that are the most useful for both the perspective of such developments and for the positive/normative issue in strict framing effects. Fairly straightforward arguments concerning the latter issue are first presented (4.2.1). Some variations in the Asian Disease's external frames that are relevant for both issues are then discussed (4.2.2). The empirical structure of the Asian Disease is finally scrutinized to show how it embeds two other types of strict framings, some features of which are non-trivial for both the theoretical and methodological purposes of this chapter (4.2.3).

4.2.1 Weak reasoned scrutiny, intelligence and incentives

One way to argue straightforwardly for the rationality or irrationality of strict framing effects is to see what happens when subjects are asked to take the time to provide a rationale for their choices. This has been done in several ways, yielding mixed results. With the original Asian Disease problem, the framing effect disappears in some experiments (e.g., Miller and Fagley 1991; Takemura 2014, pp.111-2) while remaining robust in others (e.g., Sieck and Yates 1997; LeBoeuf and Shafir 2003). There seems to be no underlying regularity about the way the experimenter (i.e., the decision modeler) asks for the rationale and the disappearance or robustness of the framing effects. For instance, this can be done by asking to briefly write the rationale (Miller

and Fagley; LeBoeuf and Shafir) *versus* asking to take some time to do it in a structured manner (Takemura; Sieck and Yates); or to write it 'publicly' *versus* 'privately' so that the experimenter respectively can or cannot read it afterward (e.g., Miller and Fagley, Takemura, LeBoeuf and Shafir *versus* Sieck and Yates, respectively). Recall that, in chapters 1 and 2, we characterized 'reasoned scrutiny' as the production and weighting of existing arguments for or against a given choice. We can say that we have here an implementation of *weak reasoned scrutiny* in so far as only the production of arguments is concerned and not the weighting of existing ones. All these studies are conducted using between-subject designs, i.e., subjects are exposed only to one of the two frames and the framing effects occurs (or not) in the modal preferences of the general population. Only Robyn LeBoeuf and Shafir use (also) a within-subject design, i.e., subjects are exposed sequentially to both frames so that framing effects can be observed (or not) for a given subject. It is arguably at this level that the positive/normative issue is most meaningful, hence some more comments on their results are in order. One specificity of LeBoeuf and Shafir's within-individual results is that they are interpreted in light of an inter-individual difference measure called 'need for cognition'. Roughly, subjects are asked, independently of their responses to framing problems, to rate a series of statements such as 'I would prefer complex to simple problems' or 'Thinking is not my idea of fun' on a scale "from -4 (very strongly disagree) to +4 (strongly agree)" (2003, p.81). The overall ratings of all these statements give individual scores that are known to be positively correlated with "more thoughtful analyses of written messages, engag[ing] in greater information search, and pay[ing] less attention to surface cues" (p.79, references omitted). Their main result is twofold. Firstly, framing *effects* (i.e., preference reversals) slightly diminish for subjects who scored high in need for cognition compared to subjects who scored low. Secondly, the influence of *framing* in the first frame presented (i.e., choosing the sure consequence in the 'save frame' and probabilistic one in the 'die frame') is the same regardless of the scores on need for cognition. In short, subjects with high need for cognition are just as susceptible to framing as those with low need for cognition, though the former slightly tend to seek more consistency by avoiding violation of description invariance.

Broadening the spirit of LeBoeuf and Shafir's experiments, another way to argue quite straightforwardly for the rationality or irrationality of strict framing effects is to look for correlation between their occurrence and subjects' *intelligence*. Indeed, cognitive psychologists and

behavioral economists have been recently (*circa* 2000) interested in the relation between the notions of rationality and intelligence. They do so by looking at correlations between inter-individual measures of cognitive ability (the need for cognition is only one among many) and departures from or abidance by standard models of individual decisions. These measures have been shown (1) to inter-correlate between one another, (2) to be positively correlated with characteristics such as success in education (and others like the ones mentioned in the previous paragraph, see Stanovich and West 2000) and (3) to be further positively correlated with so-called 'cognitive reflection tests'. The latter are simple and handy tests developed for systematic uses in experiments delivering a score which serves as a proxy for cognitive ability measure. A cognitive reflection test usually consists in answering three brief mathematical questions, e.g., "A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?" (Frederick 2005, p.26), scoring 1 point if the good answer is found or 0 otherwise. There is a fairly clear pattern of positive correlations between cognitive reflection tests' results and non-violations of normative axioms of the standard models, i.e., the lower the score the most likely the violations. However, *strict framing effects in the Asian Disease are an exception* (Toplak et al. 2014, p.159). That is, there is no significant tendency whereby the most 'intelligent' subjects are those who do not violate description invariance (contradicting the slight tendency observed by LeBoeuf and Shafir).[12]

More traditional for economists, another quite straightforward way to argue for the rationality or irrationality of strict framing effects is to look for variations in the *incentive* structure. Anton Kühberger, Michael Schulte-Mecklenbeck and Josef Perner (2002) have checked just that in a series of experiments using both real and hypothetical monetary consequences in within-subject designs (the structure of the task is akin to De Martino et al.'s (2006), though slightly more complicated). The results are quite clear: there is no difference between real *versus* hypothetical monetary consequences, but the bigger the consequences, the greater the framing effects, i.e., *small* monetary consequences make the framing effect disappears contrary to what they call "*The economists' argument*" (Kühberger et al. 2002, sect.5.1).

Summing up this subsection, three sources of normativity – weak reasoned scrutiny, intel-

[12]The arrival of the notion of intelligence in the debates around individual rationality in economics and *P*sychology, though not surprising, has (also unsurprisingly) not been uncontroversial (see Stanovich and West 2000, pp.665-701).

ligence, incentives – that have been used to argue for the rationality or irrationality of some behavioral regularities all fail to unambiguously argue for the irrationality of strict framing effects.[13]

4.2.2 Checking, inducing and varying equivalences

By contrast with the previous experiments focused on decision makers' preferences, Frisch (1993) – who coined the 'strict framing'/'loose framing' distinction – was more interested in checking their indifferences between whole decision problems (i.e., not only between consequences). This is tantamount to *checking equivalence* classes of consequences in internal frames. Doing so requires the external frames to allow for so-called "joint evaluations", i.e., both frames are presented at the same time (e.g., being presented with $\{A, B, C, D\}$ in the Asian Disease), by contrast with more traditional "separate evaluations" where they are presented sequentially in a within-subject design, e.g., first $\{A, B\}$ and then $\{C, D\}$ in the Asian Disease (see Hsee 1996 on this distinction). Frisch (1993) uses both modes of evaluations as follows. In her first experiment, subjects have to make choices in traditional (loose and strict) framing problems in separate evaluations (i.e., in both frames presented sequentially), which are then presented in joint evaluations. In the latter, there are the following instructions: "You are to read the two situations and decide whether you think the two situations should be treated the same way or whether you think they should be treated differently" with a place to write the rationale for why they thought they were different when they thought so (p.404). These rationale were then coded by two independent judges as belonging to one of three categories, to which Frisch attached different theoretical implications:

(1) *Objective reasons* if subjects inferred information that were not present in the problem, which implied that "the problems, as interpreted by the subjects, actually were different, and therefore cannot be viewed as a violation of description invariance" (p.405)

[13] At least two further results could have been added in this section: Keysar et al. (2012, experiment 1) show that the Asian Disease's framing effects disappear when performing the task in a foreign language and Reyna et al. (2014) show that intelligence agents who deal with highly risky decisions on an everyday basis are more prone to framing effects in the Asian Disease than college students and adult post-college students. However, they both argue for the irrationality of framing effects anyway. Their arguments are, roughly, because foreign language makes the problem less emotional and push one to reason more carefully than in one's native language; and because daily expertise pushes over-uses of intuitions over deliberation.

(2) *Subjective reasons* if subjects saw the difference in terms of "regret, fairness, wastefulness" (ibid) or emotional states, which implied that "subjects are violating utility theory and on reflection think it is reasonable" hence "we might question the normative status of the principle of description invariance" (ibid).

(3) *No reasons* if subjects did not give any justifications or unclear justifications; here Frisch remains silent in terms of theoretical implication).

Among the subjects who revealed the framing effect in the Asian Disease, 69% said the two situations should have been treated the same way (i.e., description invariance should not have been violated), no one stated objective reasons, 10% stated subjective reasons and 21% stated no or unclear reasons. Most other (loose) framing problems are around or (sometimes well) below 50% of those who exhibited framing effects judging the two problems to be the same after all. Hence, despite description invariance not been universally accepted as normatively sound in general, including in the Asian Disease, it is judged as more normatively sound in this problem compared with other framing problems. This conclusion will be qualified and weakened throughout the rest of this chapter by discussing other results on the Asian Disease, starting with Frisch's (1993) second experiment. The latter is the same as the first one, with the exception that subjects do not have to provide a rationale anymore, but only answer "yes" or "no" to the following question: "Do you think the difference in these two situations warrants treating them differently?" (p.410). In this case, there is a general decline in the percentages of those who exhibited the framing effect and then said the two frames should in fact be treated the same way, which is now around 50% in the Asian Disease. We shall discuss the implications of these results at the end of this subsection, but notice one methodological feature of Frisch's (1993) experiments: they show that it is possible to empirically investigate decision maker's equivalence classes, in a somewhat more refined way than Thaler's experiments on mental accounting.

Notice also that decision makers' equivalence classes are influenced by elicitation procedures (i.e., providing a rationale or not, as already seen in the previous subsection), inducing more or less match (in experiment 2 or 1, respectively) with decision modeler's equivalence classes where description invariance holds. Hidetaka Okder (2012) provides even neater procedures for *inducing equivalences* in internal frames. In his second experiment, he presents the Asian Disease

in joint evaluations with a modification of the decision scenario whereby two committees propose each one frame (one proposes $\{A, B\}$ while the other proposes $\{C, D\}$). There are four decision questions: (1) "From committee 1, do you prefer Program A or B?", (2) "From committee 2, do you prefer Program C or D?" (3), "Do you think that Program A in committee 1 and Program C in committee 2 are equivalent?" and (4) "Do you think that Program B in committee 1 and Program D in committee 2 are equivalent?". One result is that the modal preference reversal disappears as 77.1% of decision makers have consistent preferences here (the modal pattern is slightly for the sure consequence, i.e., $[A\&C]$). However, the great majority of the remaining 22.9% do not think that there is an equivalence between the two frames (i.e., they answered "no" in both (3) and (4)). Notice that the *choices* are made in joint evaluations, by contrast with Frisch's external frames where only the elicitation of the equivalence classes was made in joint evaluations. In his third experiments, he presents the Asian Disease in separate evaluations, but adds six decision questions before the original one asking which of the programs is favored, e.g., for the 'save frame' $\{A, D\}$ (the italicized beginning is common to all six questions): "*How many people will be estimated to be* (1) saved when no program is adopted? (2) lost when no program is adopted? (3) saved when Program A is adopted? (4) lost when Program A is adopted? (5) saved when Program B is adopted? (6) lost when Program B is adopted?". One result is that the modal preference reversal disappears even further than in the previous experiment, i.e., nearly everybody is consistent (again the modal pattern is slightly for the sure consequence) and provide answers in (1)-(6) matching with the good answers from the decision modeler's perspective, i.e., 0 to (1), 600 to (2), 200 to (3), 400 to (4) etc.

In a way, Okder partly answers a rather straightforward and intuitive criticism of the Asian Disease, which is made explicitly by Wakker: "The message "200 people will be saved" does not make clear what will happen to the other 400 people, whether they will die or not"[14]. But it would be a mistake to conclude from Okder's results that when there is no more ambiguity the Asian Disease is fully explained and there are no paradoxes anymore (because no more preference reversals and framing effects). This would beg the general methodological implications arising from the literature on framing effects taken as a whole: how is the remaining information to

[14]From the comments on Tversky and Kahneman (1981) in Wakker's annotated bibliography available at http://people.few.eur.nl/wakker/refs/webrfrncs.docx, last consultation: 21/01/2016.

described (or 'framed') in ordinary language? The following results illustrate how preferences are not invariant *also* to the different ways by which descriptions of the consequences in the Asian Disease can be made more or less complete (all the following illustrations have decision scenarios and questions identical to the original Asian Disease). For instance, completing the sure consequences by descriptions from their respective '*save*' and '*die*' external frames restores consistency:[15]

> If Program A' is adopted, 200 people will be saved and 400 people will not be saved.
> If **Program B** is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.
> If Program C' is adopted, 400 people will die and 200 people will not die.
> If **Program D** is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. (Kühberger 1995, exp.1)

But using *only* these complementary descriptions triggers a framing effect which is the reverse of the original one:

> If Program A'' is adopted, 400 people will not be saved.
> If **Program B** is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.
> If **Program C**$''$ is adopted, 200 people will not die.
> If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. (Kühberger 1995, exp.2)

Finally, mixing the original descriptions from the save and die frame pushes the modal preference towards the sure consequence:

> If **Program A**$'''$ is adopted, 200 people will be saved and 400 people will die
> If Program B''' is adopted, there is 1/3 probability that 600 people will be saved and 2/3 probability that 600 people will die. (Mellers and Locke 2007, p.362)

Varying equivalences between frames in a strict framing problem is essentially a linguistic exercise. The above variations played on two basic linguistic devices to generate arguably equivalent statements: presence or absence of negations ('not') and opposites ('saved'/'die'). Bringing synonyms into the mix (e.g., 'survived'/'perished') let us appreciate how the possibility of different descriptions for the same consequences and combinations by pairs to constitute Asian Disease

[15] A similar manipulation tend to cancel intertemporal preference reversals, i.e., making explicit that if the smaller consequence is chosen now then *no consequence will occur in the future* and if the later consequence is chosen then *no consequence occurs now* (see Magen et al. 2008). I am not aware of similar manipulations for social preferences.

frames may well be *infinite*. As preferences are already influenced in all directions within the small subset of this infinity considered so far, it seems rather misplaced to hope for clear-cut empirical patterns. In my opinion, the only empirical fact worth considering for a formal account of strict framings of consequences is just the observation made in this paragraph, namely that all consequences can be described in an infinity of manners by virtue of presence or absence of negations, opposites, synonyms (which only constitute some of the most basic linguistic devices to generate different descriptions of the same thing), and that preferences can exhibit all sorts of patterns within this infinity. Trying to formally capture the details of the underlying extensive power of ordinary language uses is hard enough for linguists.[16]

Summing up this subsection, there is a variety of means to construct external frames that allow the study of how equivalences within internal frames can be empirically checked, induced and variated. It is quite obvious that the variations cannot be straightforwardly predicted and explained by prospect theory, or by any existing theory for that matter (for detailed statements of this claim, see Kühberger 1997; and Takemura 2014, chap. 10). Furthermore, it is not the case that all violations of description invariance implied by these variations are always deemed irrational by the decision makers who commit them, though of course some do consider it as irrational and would reverse their preferences. The main lesson from this subsection is that a formal account of strict framing phenomena should allow for the logical possibility of representing the infinity of possible different descriptions of same consequences, without imposing *a priori* a peculiar pattern of preferences. It seems wiser to account for such patterns parametrically, but what features of the decision situation should be parametrized remains unclear so far. The next subsection along with the next section may suggest some clues to clarify this issue.

4.2.3 Strict framings of attributes and goals within the Asian Disease

An influential typology of framing effects proposed by Irwin Levin, Sandra Schneider and Gary Gaeth (1998) classifies the Asian Disease as belonging to the "risky choice framing effect" subset of a "valence framing" set, which contains two other subsets, the so-called "attribute framing"

[16] Further experiments varying equivalences in the Asian Disease are summarized by Kühberger and Tanner (2010, Table 2; see also Schulte-Mecklenbeck and Kühberger 2014); Tombu and Mandel (2015) provide other ones with financial problems, along with further references varying equivalences in strict framings of consequences (see also Chick et al. 2015).

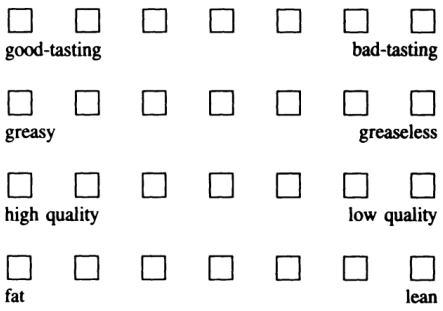

Figure 4.2: Decision questions from Levin (1987)

and "goal framing". The paradigmatic example of an attribute framing is Levin's (1987) ground beef example where two frames, a 'lean frame' and a 'fat frame' are used to describe the same ground beef as follows (p.85):

> [decision scenario:] In each pair we want you to indicate by filling in one of the squares which item in the pair you are most apt to associate with a purchase of 75% lean ground beef and the extent to which you associate the purchase with that item rather than the other item in the pair.
> [decision questions:]
> See Figure 4.2.
> Results: relative tendency toward **good-tasting**, **greasless**, **high quality** and **lean**
>
> [non-identical decision scenario:] In each pair we want you to indicate by filling in one of the squares which item in the pair you are most apt to associate with a purchase of 25% fat ground beef and the extent to which you associate the purchase with that item rather than the other item in the pair.
> [identical decision questions, i.e., identical set of pairs]
> Results: relative tendency toward **bad-tasting**, **greasy**, **low quality** and **fat**

The paradigmatic example of goal framing is Beth Meyerowitz and Shelly Chaiken's (1987) Breast Self-Examination [BSE] example where two frames, an 'inaction leading to bad consequences frame' and an 'action leading to good consequences frame', are used to describe the results of the same medical researches on BSE (p.504; only the crucial passage of the scenario is reproduced here):

> [decision scenario:] "Research shows that women who do not do BSE have a decreased chance of finding a tumor in the early, more treatable stage of the disease"
> [decision questions (not reported by the authors) on "9-point scales" attitudes and intentions, and "5-point scales" behaviors]
> Results: relatively **strong** intentions, attitudes and behaviors toward BSE
>
> [non-identical decision scenario:] "Research shows that women who do BSE have an increased chance of finding a tumor in the early, more treatable stage of the disease."

[identical decision questions]
Results: relatively **weak** intentions, attitudes and behaviors toward BSE

The main similarity with the Asian Disease is that both attributes and goal framing are strict framings. Both involve pairs of frames that only differs by their different descriptions of identical consequences, of the ground beef to be consumed and of the BSE to be performed. Hence they show further impacts of ordinary language uses on decision makers. One minor dissimilarity is that the framing effects here are obtained by keeping the objects of choice (i.e., the scales within the decision questions) identical across frames but changing the descriptions in the decision scenario (as in the *loose* isolation effect), while the reverse is the case in the Asian Disease. A more important dissimilarity from the perspective of economics is that attribute and goal framings do not involve, at least in the experimental setup from the study referenced above, *choices* strictly speaking (though 'behaviors' in the BSE refers to whether or not subjects actually performed it some times after the experiments). But it should not be complicated either to replace the ratings scale by a willingness to pay question, or a yes/no question (or to introduce appropriate objects of choice, e.g, the 75% lean framed ground beef *versus* another type of meat, and the 25% fat framed ground beef *versus* the same other type of meat). We would thus have another strict framing of consequences under risk (BSE) and under certainty (ground beef), and if the results from the scales carry over, we would thus have strict framing effects.[17]

By contrast with Levin et al.'s arguments for distinct theoretical treatments of the internal frames in risky choice, attribute and goal framing effects, I want to point out the empirical embeddedness of their external frames with its implications for understanding the Asian Disease from the perspective of this dissertation. Let's first focus on attribute framings within the original Asian Disease. The framings of consequences in the Asian Disease are indeed nothing but framings – as "will be saved" under one frame or "will die" under another – of the attributes of the set 600 people and its subsets including the empty set . Within the objects of choice, there are six such framings of attributes, i.e., three by frames: (1) the subset 200 people *will be saved* in A, (2) the subset 400 people *will die* in C, the set 600 people (3) *will be saved* in B or (4) *will die* in D and the empty set zero people in (5) *will be saved* in B or *will die* in D.

[17] See Kreiner and Gamliel (2016) and the references therein for manipulations varying equivalences in attribute framings. Unsurprisingly, presenting a complete and mix description of the attribute reduces the framing effect. I am not aware of similar manipulations in goal framings.

Within the decision scenario, there is one such framing of attributes, namely the set 600 people is "expected to [be] kill[ed]", i.e., *will die*. Observing that framings of attributes loom large in the Asian Disease matters especially because of a result established within researches on attribute framings, namely that the effect is not only on the ratings of the objects of choice but also on the actual experience of its consumption. The classical demonstration of this is from Levin and Gaeth (1988). They extended the above results on ground beef in two further experimental conditions: one where subjects *actually consumed* 75% lean-25% fat ground beef *before the ratings*, and another one where they did so after the ratings. That the framing effects, though slightly reduced, still held strongly suggests that different descriptions of the same ground beef influence the utility derived from its experienced taste. Intuitively, it can be argued that this is not a very surprising result. Think of fancy restaurants and customers spending so much efforts and money to respectively make and consume well presented food, with sometimes very weird and long descriptions on the menus.[18]

Such reasoning in terms of utility can be translated in terms of disutility in the Asian Disease. The negative mental experience associated with the choice of a program that will lead to the death of some people will be smaller when these deaths are left implicit than when they are made explicit in the descriptions of the programs. If you had to choose just between A and C, i.e., just between two descriptions of the same consequences, it is not totally inconceivable that the disutility derived from C would be larger than the disutility derived from A. This observation is especially relevant regarding the key role of the sure consequences in the Asian Disease as discussed in the previous, the framing of which has the exact same structure as traditional framings of attributes (i.e., left the proportion of the opposite attribute implicit). There is some indirect evidence for this reasoning coming from Marc van Buiten and Keren (2009, experiment 3). They presented the two frames of the Asian Disease in a joint evaluation mode with the equivalence between frame imposed explicitly by both an arithmetical justification and an explanation that the two programs A (i.e., A and C of the original version) and two programs B (i.e., B and C of the original version) were redescriptions of the same programs. Since the

[18]For further non-conventional evidence supporting this intuitive argument, see the story that made the news in October 2014 about McDonald's food being untruthfully presented as 'organic food' to food experts who for tests, most of whom declared that it tasted way better than McDonald's, e.g., http://www.dailymail.co.uk/news/article-2804875/Dutch-pranksters-fool-food-experts-organic-convention-believing-MCDONALD-S-actually-new-healthy-dish.html, last consulted 23/01/2016.

four descriptions were part of the decision scenario, the objects of choice were just A and B without descriptions. The results are an equal distribution of preferences between A and B, but to the question "which formulation do you consider most convincing as a reason for your choice?" there is a very strong modal influence of the 'save frame', i.e., regardless of the actual choice for A or B. Much more direct evidence for the reasoning presented above comes from Kühberger and Patricia Gradl's (2013, experiments 3 and 4). Subjects were presented the Asian Disease in a joint evaluation, and had to rate *all four programs* on three scales: (1) likeliness to choose, (2) believed efficacy and (3) emotional impact. The modal preferences from (1) are $A \succ B \sim D \succ C$, which highly correlate with the efficacy ratings from (2), both of which weakly correlate with the emotional ratings from (3) in the overall, though the correlation is much stronger when only the emotional ratings for the sure consequences are considered (experiment 3). The modal preferences $A \succ B \sim D \succ C$ is also found in their last experiments where subjects had to rank directly (i.e., no ratings anymore) the four programs (experiment 4).[19]

We can now turn to the framings of goal within the Asian Disease, which is a slightly trickier issue. Within the decision scenario, the decision modeler expresses his goal to "combat the disease". That is, the goal is to do something, which is easily shared by the decision maker, by contrast with not doing anything and letting 600 people die. Within the decision question, the decision modeler tries to elicit the decision maker's goal. That is, the latter has to express what he will do (choose) with respect to what he "favor[s]" (prefers), by contrast with either the opportunity of not choosing anything or the opportunity of expressing his goal with respect to what he disprefers by selecting the program he does *not* choose which is also known to create framing effects (Shafir 1993). Hence, framings of goals within the decision scenario and decision question are quite obvious and identical across frames. Whether there are framings of goal within the objects of choice could be a controversial issue. Under one interpretation there are no framings of goals, period. This is the most likely to correspond to the decision modeler's interpretation if he is an experimenter working in decision theory or behavioral decision theory

[19]Note that, this partly undermines Kahneman's (2011) interpretation – through De Martino's (2006) neurobiological results – of the Asian Disease framing effect as underlying emotional processes evoked by the wording of the problem, because it is not clear why these emotions should be evoked *only* from the words of the sure consequence but not from the exact same words in the probabilistic ones. For further neurobiological results on strict framings of consequences undermining Kahneman's (2011) interpretation in this direction, see Yu and Zhang (2014, esp. p.8). Also, for a formally articulated philosophical explanation of Kühberger and Gradl's results, see Geurts (2013).

where individual decisions are conceived as taking place between a purposeful decision maker and the purposeless 'Nature'. Under a second (opposite) interpretation, there are framings of goals because the ones in the decision scenario and question somewhat 'leak in' the objects of choice. With respect to the decision modeler's goal "to combat the disease", that people are "saved" (Program A) is *congruent* with that goal, while that they "die" (in Program C) is *incongruent*. Restricting our attention to A and C, which is justifiable from the key role they play in the framing effect as discussed in the previous section, the difference across frames can be interpreted as an asymmetry of congruence with respect to the decision modeler's goal. Hence this second interpretation provides further argument against the equivalence of frames in strict framings. Under a third and fourth (in-between) interpretations, there are either neutral framings of goal (third interpretation) *or* random framings of goals (fourth interpretation) because of the mix in the semantics of the probabilistic programs arising from the occurrence and non-occurrence of *negations* in "*no*body" and "*no* people". The set of work discussed in the next section provides arguments to characterize the conditions under which the second interpretation is more likely to be true.

Summing up this subsection, considering the embeddedness of risky choice, attribute and goal framings, along with further results established in the respective literatures, suggests two reasons why the two frames of the original Asian Disease would not be considered as equivalent. On the one hand, they might not be equivalent because parts of the utility derived during the experience or realization of the consequences come from the descriptions of their attributes. On the other hand, they might not be equivalent because of an asymmetry between the expression of the decision modeler's goal to save people (which is arguably easily shared by the decision maker) and the expression of the decision maker's goal about how he actually intends to implement it with respect to his preferences.

Conclusion

This section summarized a set of arguments based on experimental results for the non-systematic imposition of description invariance in models of individual behaviors, for both positive and normative reasons. With respect to the conclusion of the previous section, it should be noted, from an historical perspective, that the independence axiom and the axiom of description invariance

were both initially hidden in the formalism of the standard model. Both were made conceptually and formally explicit only *ex post* (see Malinvaud 1952; and Fishburn and Wakker 1995 on the independence axiom). This calls for one further comment on the work of Frisch (1993). Her experiments were explicitly inspired by the ones pioneered by MacCrimmon discussed in chapter 2 that led to theoretical developments weakening the axiom of independence. Her experiments were also inspired by Levin and Gaeth's (1988) results on the increased experienced utility from descriptions along with Kahneman and Tversky's comments on the rationality and irrationality of framing effects mentioned in the previous section. Frisch argued that, against this background, her results warranted a weakening of the axiom of description invariance just as the independence axiom had been weakened. The arguments made in this section and in the next one are meant to motivate this project further, at least for strict framing phenomena.[20]

4.3 The communicative structure of external and internal frames

There is a growing trend in *P*sychology that studies strict framing phenomena through the communicative interaction between the decision modeler and the decision maker, though they are respectively called the *speaker* and the *listener* or *hearer*. For the sake of giving a broadly coherent account of framing phenomena in this chapter, we shall discuss this trend in a way that develops the 'goal leaking' interpretation suggested at the end of the previous section. We shall see how the decision modeler's choice of one description over another may leak his goal (4.3.1). Then, how this leaking can be understood more precisely by delving into the descriptive structure of a consequence (4.3.2). And finally how the philosophical and linguistic backgrounds of these contributions are consistent with both the one from this dissertation in terms of speech act theory (cf. chapter 1) and decision theory in economics (4.3.3).[21]

[20] Though Frisch's position was rather uncommon at the time (compare to now, see the next section), an even earlier similar position was held by philosopher MacLean (1985).

[21] Most of contributors to this trend are reunited in *Perspectives on Framing* (Keren 2011b). Besides them, two prominent contributors on framing in *P*sychology, David Mandel and Kühberger, have lately expressed strong sympathy with this trend, see Mandel and Vartanian (2012), Mandel (2014; 2015), Tombu and Mandel (2015), Kühberger and Tanner (2010), Kühberger and Gradl (2013).

4.3.1 Goal leaking through the choice of a description

A crucial motivator of the aforementioned trend is the *information leaking* account of framing (and other) phenomena advanced by Craig McKenzie with various co-authors (especially Shlomi Sher). The following is a vivid illustration of an experimental demonstration of how the mere *choice* of one description over another leaks choice-relevant information for the decision maker – here the information is about a reference point and the demonstration works by symmetrical pairs of experiments with hypothetical scenarios (McKenzie and Nelson 2003, experiments 1 and 3). In the first experiment, subjects have a clear reference point: they are asked to imagine being in front of a 4-ounce cup which is either *full* in one frame, or *empty* in the other frame. They briefly leave the room and, *unobserved by them*, some changes occur to the contents of the glass: when they come back there is, irrespective of the frame, water at the 2 ounces level. In the other experiment, subjects are told a similar story about Mary: she is in front of a glass, the contents of which is *unknown to the subjects*. Mary briefly leaves the room, and, seeing a change in the contents of the glass when she is back, says either "The glass is half full" in one frame or "The glass is half empty" in the other frame. The decision questions are symmetrical. In the first experiment, subjects have to choose the most natural of two descriptions: "The glass is half full" or "The glass is half empty". And in the second experiment, they have to infer the most likely of two states: "The glass was full before its contents changed" or "The glass was empty before its contents changed". There is the following symmetrical framing effects in both experiments. The modal preference of those starting with an empty cup is for the consequence of the changes described as a half-full cup (and *vice-versa* for those starting with a full cup). Hence *different initial reference points induced different descriptions of the same consequence.* And symmetrically, *different descriptions of the same consequence by Mary induced different inferences about her initial reference point* because those to whom she said the glass is half-empty made the modal inference that the glass was full. Taken together, these results suggest that, by virtue of linguistic regularities underlying a shared language, one's *choice* of description can leak information – here about the initial reference point – that goes beyond the description itself, and this information can be rationally inferred by the one to whom the description is

made.[22]

Two features of the previous experiments are worth noting. First, the framing effects are *loose* ones and the inference is precisely about what makes the two situations non-identical, i.e., a glass that is initially full *or* empty, by contrast with conventional uses of 'the glass half-full or half-empty' to express difference of perspectives on the same situation. Secondly, the inference is about states of affairs in the world, and *not* about, though *drawn from*, what someone said and her potentially related beliefs, preferences or other mental states, e.g., the contrast between optimism and pessimism conventionally expressed by 'the glass half-full or half-empty' is not relevant here (at least not 'directly' relevant). Both features are different in Sher and McKenzie's (2008, pp.89-90) experimental investigation of the Asian Disease. They used their "embedded creativity" experimental design (2006), where subjects have to construct (i.e., 'create') a frame or a description, rather than choosing one that is fully already made (as were the descriptions in the previous paragraph). After subjects are instructed about the structure of the two programs in the Asian Disease, and have expressed their own preferences for either one of the two programs (in a mixed frame, i.e., both 'saved' and 'die' within one description; personal communication with Sher), its embedded creativity version runs as follow (2008, p.90):

> "Imagine that your job is to describe the situation [...] to a committee who will then decide which program, A or B, to use. Please complete the sentences below as if you were describing the programs to the committee"

The sentences to be completed are reproduced in Figure 4.3.

The main results, the preliminary nature of which is emphasized by Sher and McKenzie (2008), are: (1) among those who prefer the probabilistic program there is roughly an equal distribution between framing the sure one as '200 saved' and as '400 die', (2) regardless of prior preference there is a modal framing of the probabilistic program as '1/3 of 600 saved and 2/3 of 600 die', and (3) among those who prefer the sure one there is a strong modal framing of the sure one as '200 saved'. As '200 saved' corresponds to the modal preference of decision

[22]Sher and McKenzie (2006) replicated McKenzie and Nelson's (2003) results with non-hypothetical decision scenarios, e.g., with tasks involving real glasses of water. It should also be noted that McKenzie and Nelson's (2003) results and theoretical interpretations are more complex than discussed in this paragraph. Besides including other quantities (e.g., 1/4 empty 3/4 full), they hypothesized and observe a tendency to describe a consequence by focusing on what has *increased* with respect to the initial reference point. As will be argued later, this is the kind of empirical subtleties that formal account of framing effect in economics should allow for, without systematically implying it, at least at the axiomatic level (i.e., not parametrically in a given model).

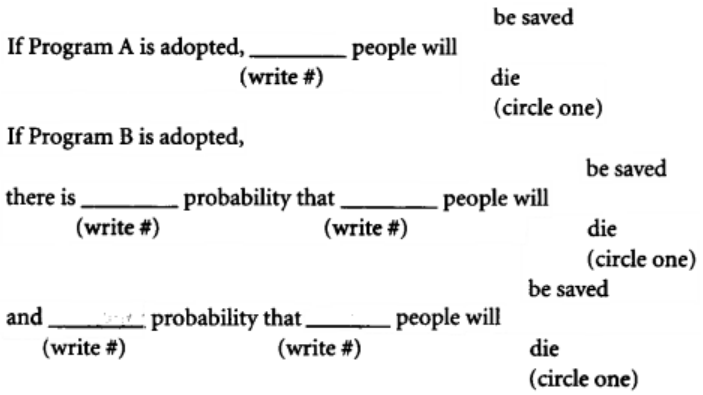

Figure 4.3: Sher and McKenzie's (2008) embedded creativity for the Asian Disease

makers in the original version and in general (cf. previous section), (3) suggests one non-trivial informational leaking in the Asian Disease, namely about the decision modeler's own preferences leaked through the saved framing of the sure program.

Sher and McKenzie (2008) argue that if this informational leak is interpreted by the decision maker as an implicit recommendation, then the original preference reversal is not necessarily irrational. The conditions under which this argument is most likely to hold, at least in the Asian Disease, have been investigated by van Buiten and Keren (2009, experiments 1 and 2). These conditions bear on what they argue is a joint/separate evaluation modes asymmetry between decision modeler and the decision maker. While the former usually chooses an external frame in a joint evaluation mode (i.e., $\{\{A, B\}, \{C, D\}\}$), the latter's internal frame is usually constructed from the external frame presented by the decision modeler in a separate evaluation mode (i.e., either $\{A, B\}$ or $\{C, D\}$). The idea is that there is information leakage when the decision modeler puts himself into the decision maker's shoes when choosing the external frame. This reasoning is experimentally translated by instructing subjects to promote one of the two programs and that they have to choose one of the two original frames in which the decision maker will then make his choice. When the two frames are presented in a joint evaluation mode, the modal choice is the 'save frame' regardless of the program that should be recommended (experiment 1). But when the two frames are presented in a separate evaluation mode, then the modal choices are the 'save frame' for those who have to promote the sure program and the 'die frame' for those who have to promote the probabilistic program (experiment 2). These results suggest

the following refinement to Sher and McKenzie's (2008): information leaking about the decision modeler's preference for the sure program is likely to be non-intentional if the latter constructed the external frame without putting himself into the shoes of the decision maker, i.e., in a separate evaluation mode where the impact of linguistic regularities underlying a shared language becomes more salient. When decision modeler are in the latter mode of evaluation, intentional information leakage about all programs is more likely, though not necessarily about the decision modeler's preferences and not necessarily designed to be inferred by the decision maker, e.g., about a recommendation imposed to the decision modeler (against his own preference or not) and designed to influence decision makers (against their own preferences, e.g., manipulation, or not, e.g., benevolent nudging).[23]

Arguably, the account of the strict framing of consequences in the Asian Disease given so far in this subsection is in line with the interpretation from the previous section whereby decision modeler's goal can leak in the objects of choice through the description of consequences. To close this subsection, I propose to speculate on three possible and non-mutually exclusive ways to refine this interpretation, all inspired from other areas of research on strict framing effects.

Refinements from the literature on goal framing from *P*sychology are less straightforward than one would thought because of a lack of investigation of the communicative structure of choices in the same fashion as the contributions discussed in this subsection (see Rothman and Updegraff 2011). The notion of goal in this literature has been mostly developed by applying the so-called *self-regulatory focus* framework on decision makers' internal frames (see Higgins 2012). Roughly, in the latter, (1) the construction of goals, (2) choices to attain them and (3) related experiences of consequences within internal frames are *promotion focus* when guided by the ideal of getting the optimal positive consequences and *prevention focus* when guided by the obligation of avoiding the maximal negative consequences. Kühberger and Christian Weiner (2012) have investigated strict framings of consequences in a monetary version of the Asian Disease under this framework. Their main results are that the correlation between kinds of regulatory focus

[23]Two points are worth mentioning. First, McKenzie et al. (2006) discuss finer explanations and results on implicit recommendation, but in the domain of default options (e.g., about how the choice of a default option can leak, in a non-mutually exclusive fashion, information about the decision modeler's own preferences, or beliefs about what people ought to do, or beliefs about what people would like to do; and how these information are rationally inferred or not). Second, van Buiten and Keren (2009, experiment 1) hypothesized and observed a positivity bias documented elsewhere in *P*sychology, explaining the tendency to select the positive frame in joint evaluation mode irrespective of the imposed program to be promoted. Again, a formal account of framing effect in economics should, at the axiomatic level, allow for this kind of empirical regularities without implying them.

and gain-loss frames is at best very weak, and that, *irrespective of frames*, clearly prevention focused decision makers strongly tend to choose the sure program and clearly promotion focused decision makers strongly tend to choose the risky program. Hence there is a strong effect of the congruence of decision makers' goals with consequences *per se* but not with their framings. As Kühberger and Weiner emphasize, their results open a lot of questions, to which we can add two further sets of questions motivated by the previous section. The first set concerns the decision makers' goals. Their results were obtain in between-subjects design. If regulation focus has strong effects independently of framing then framing effects should disappear in within subject-subjects design, which is usually not the case. How can that be? Furthermore, their results were obtained in the classic version of the framing of consequences (e.g., '€200 will be saved' versus '€400 will be lost'). Would they hold in the variety of other framings were framing effects usually disappear or reverse? The second set concerns the decision modelers' goals. As argued in the previous section, the implicit goals within the descriptions of the consequences stem from an interplay with the description of the decision scenario and question. It seems possible to manipulate the regulatory focus of the decision modeler by making it more explicit. For instance, compared with the original consequences, the original "is expected to kill 600 people" and "to combat the disease" in the decision scenario seem either without clear regulatory foci or *mildy* prevention focus. '600 people are expected not to live because of an unusual Asian Disease' and 'to save these people' would seem more clearly promotion focus. Would such manipulation trigger an effect of congruence between decision makers' and decision modeler's goals?[24]

Further refinements can be made from the literature on so-called 'direction of comparison effects'. Classic direction of comparison effects involve contradictory answers to pairs of questions such as 'How similar is China to North Korea?/How similar is North Korea to China' or 'Does traffic contributes more to air pollution than industry?/Does industry contributes more to air pollution than traffic?'. As discussed by Michaela Wänke and Leonie Reutner (2011), direction of comparison are quite robust and the main feature that matters is not the order of words per se but which entity is put as the subject (to be evaluated) and which is put as the referent

[24]Two points are worth mentioning. First, the prevention/promotion focus distinction seems to be currently under debate regarding its status as an interindividual difference grounded in the decision makers which would be 'chronically' of either kinds *or* in decision situations which would contingently 'induce' either kinds. Second, Kühberger and Weiner measured types of self-regulatory foci either by verbal protocols analysis of decision makers instructed to think aloud when making their choices (experiment 1) or by questionnaires (experiment 2).

(providing the norm of evaluation). For instance, in 'How similar China is to North Korea?', China is the subject and North Korea the reference, hence diplomatic issues with the United States or nuclear issues are more likely to come up as an evaluative dimension than with the reversed question. Wänke and Reutner (2011, p.187) argue that exactly what it means to be the subject in a sentence needs to be investigated to fully understand direction of comparison effects. On their account, subjects "indicate the narrative theme or topic", providing *cues* for the reader or listener about the "conversational goal", and allowing them to infer what is of interest to the speaker in order to make their eventual answer to his eventual question relevant to him (ibid). Translated to the Asian Disease, it suggests that some effects may be observed if the conversational goal imposed by the decision modeler would change by taking the people at stakes as subjects and the disease and programs as referent, e.g., '600 people are expected to die from an unusual Asian Disease...', '200 people will be saved by implementing Program A'. As far as I know, it is exactly the reverse in virtually all versions of the Asian Disease used in the literature.

Finally, refinements could be made from the so-called 'constructive processing' perspective on framing effects. The constructive processing perspective on framing effects stems from the very elegant results on the Asian Disease obtained by Herbert Bless, Tilmann Betsch and Axel Franzen (1998). The latter simply added a simple "context cues" (ibid) to the original Asian Disease: in one condition 'statistical research' was printed in a corner of the page on which subject responded to one frame, and 'medical research' was printed in the condition. There was still a framing effect with the medical cue but it disappeared with the statistical cue. As discussed by Eric Igou (2011), this result is quite robust and has been interpreted in terms of the need for the decision maker to infer further information than the one given in the medical (but not statistical) problem. On his account, these further inferences are made through a constructive processes whereby frames become "affective cues" (p.222) guiding the construction of the decision maker's goal. Hence the interaction between decision modelers' and decision makers' goals seem highly moderated by the interaction between the contextual cues provided by the former, inducing the frames to be affective cues for the latter.

I would like to sum up this subsection under the analysis of Sher and McKenzie (2006) on attribute framing. The results discussed in this subsection (and to a lesser extent in the previ-

ous section) show that description invariance, as usually conceived between pairs of logically or mathematically equivalent problems, does not necessarily imply *information invariance*, especially the invariance of *choice-relevant information*. While two descriptions may indeed provide equivalent choice-relevant information, the choice of one of these description over the other by the decision modeler is itself a relevant information for the decision maker's choice. The decision modeler's choice of description *leaks* information that makes the *informational content* of two descriptions of the same consequence non-equivalent for the decision maker. What exactly is this leaked information about seems a highly situation-dependent matter, contingent upon subtle details of the linguistic interaction between the decision modeler and the decision maker. If "psychologically salient properties recruit congruent linguistic terms", then there is a potential systematicity in the decision modeler's goal to be signaled (or 'cued') by his choice of description, hence informational leaking cannot be dismissed as random noise (Sher and McKenzie 2006, p.488). If the decision maker has a "sensitivity to subtle linguistic cues" (ibid, p.488), then he absorbs this information. Exactly how remains unclear. It does not seem to be a classic inductive inference because creativity and imagination (possibly through narrativity) are involved in the process of going beyond literal information. But at the same time it seems to be an unconscious process, "[o]therwise, the non-equivalence of attribute frames would have been self-evident prior to our analysis, and no disturbing conclusions about human rationality would have been drawn from attribute framing effects" (p.489). Following Sher and McKenzie (2006, pp.491), it can be argued that "[strict] framing effects are best understood, not as paradoxes of rationality, but as paradoxes of measurement": "framing effects *without information equivalence* raise the question of whether the analysis of preferences is being undertaken at sufficiently high resolution" (p.492). From the perspective of giving a formal account of framing phenomena in economics, two levels can be distinguished. At the axiomatic level, it seems reasonable only to increase the resolution of the analysis a little bit, just to take account of the fact that different descriptions of the same consequence can have different informational contents. This is the level at which the formal contribution in the next chapter will be situated. At the modeling level, getting acquainted further down in the linguistic dimension of informational leakage may be useful to make quantitative predictions. The next subsection discusses a set of potential inspiration for this project, along with its potential limits.

4.3.2 Goal leaking within the minimal descriptive structure of a consequence

If going from consequences to their different descriptions increases the resolution of analysis, then going from these descriptions to their "independently moving parts" (Sher and McKenzie 2011, p.45) would further increase the resolution of analysis. It can be argued that *the minimal descriptive structure of a consequence* is constituted by four (for choices under certainty) or five (for choices under risk) types of such independently moving parts, expressing respectively:

(1) a quantity, e.g., "200", "400";

(2) what is being quantified, e.g., "people";

(3) an attribute of what is being quantified, e.g., "saved", "die";

(4) the presence or absence of negation, which, at least in the Asian Disease, are used to negate the attribute, e.g., "*not* be saved", "*not* die", or not used to express it, e.g., "be saved", "die";

(5) the probability of occurrence of one combination of (1)-(4), e.g., "1/3 probability that", "2/3 probability that".

Different descriptions of a same consequence can be generated by appropriate combinations of one instance from each of the first four types, including one from the last type for consequences under risk. Insights about the framings of each of these types have been gathered by psychologists over the years, most of whom are contributors to *Perspectives on Framing* (Keren 2011b).

Denis Hilton (2011) provides the most systematic discussion of these insights by focusing on the pervasive *polarity* that underlies the uses of logical expressions in ordinary language. Two types of such logical expressions studied by psychologists are the so-called 'natural language quantifiers', i.e., the ordinary language translation of numerical information, about either (1) or (5). To illustrate with (1), consider the contrast between '*a few* people will be saved' and '*few* people will be saved'. For a given situation, even if 'a few' and 'few' denote the same quantity, the connotations they convey about that quantity is nevertheless usually of opposite polarity.

In the Asian Disease situation, 'a few people will be saved' is positively connoted by contrast with 'few people will be saved' which is negatively connoted. But *vice-versa* in another situation where the goal would be to kill as much people as possible through the same Asian Disease: 'few people will be saved' would be positively connoted and 'a few people will be saved' would be negatively connoted. To illustrate with (5), the same reasoning and inversion would occur in the contrast between 'it is *possible* that everybody will be saved' and 'it is *uncertain* that everybody will be saved' (i.e., 'possible' and 'uncertain' can denote the same probability but have an opposite polarity of connotation). Though these considerations do not bear directly on the Asian Disease because of its use of numerical information (e.g., '200', '1/3'), the underlying regularities in everyday language uses have indirect implications.[25]

Quite intuitively, works on the effects of natural language quantifiers on decision makers suggest that positive polarity tends to trigger reasons *for* the proposition expressed (e.g., a program to combat or instill a disease) while negative polarity tends to trigger reasons *against* the proposition expressed (see Moxey 2011; Teigen 2011). Hilton (2011) argues that because of the pervasive polarity of logical expressions in ordinary language uses, the decision modeler cannot be neutral in communicating such propositions. That is, regardless of issues of intentionality, consciousness, persuasiveness and the like, information leakage is *unavoidable* because of this pervasive polarity. Indeed, polarity is also pervasive in combinations of (3) and (4). Linda Moxey (2011, pp.131-133) argues that the polarity underlying (3) in the Asian Disease, i.e., 'saved' *versus* 'die', can be explained straightforwardly through the extension of the results on natural language quantifiers. The positive polarity of 'saved' generates reasons for the sure program, the negative polarity of 'die' generates reason against the sure program. This further extends by inversion to (4): the negative polarity of 'not be saved' and the positive polarity of 'not die' respectively generate reasons against and for the sure program. This provides a straightforward explanation to the variations on framings presented in the previous section (from Kühberger 1995), including the absence of framing effects when both polarities are present within '200 will be saved and 400 people will not be saved' and '400 people will die and 200 people will not die'. Notice, however, that this does not explain the return of a modal preference

[25] Using numerical expressions may be *slightly* more neutral than using natural language quantifiers (as argued for instance by Keren 2012).

for the sure program when both polarities are present within '200 people will be saved and 400 people will die' (from Mellers and Locke 2007, p.362).

Indeed, as discussed by Yaacov Schul (2011) negations play a very tricky role in everyday uses of ordinary language, which is far more complex than just 'meaning the opposite'. Schul distinguishes *bipolar* attributes such as 'saved' and 'die', the negations of which brings the opposite to mind, i.e.., 'not saved' brings 'die' to mind and 'not die' brings 'saved' to mind, from unipolar attributes such as 'romantic' which does not have a clear opposite which can be brought to mind by 'not romantic'. There is an intimate relation between (3) and (4) within the minimal descriptive structure of a consequence because strict framing phenomena work by pairs of decision problems involving bipolar attributes. Hence, the question is: what lies behind the choice of describing an attribute as the opposite of a negation (e.g., 'die' instead of 'not saved', 'saved' instead of 'not die') *versus* as the negation itself (e.g., 'not saved' instead of 'die', 'not die' instead of 'saved')? One possibility is the communication of a norm or deviation from a norm through what linguists and psycholinguists called *markedness*. Pairs of opposite terms such as bipolar attributes usually carry a contrast between one of the two terms being the *unmarked* in expressing "the usual, the normal, the common, and the neutral or less specific" (p.165) by contrast with the *marked* one expressing something more peculiar, less usual etc. Suppose for a moment that 'die' is the marked term and 'saved' the unmarked one in the 'die'/'saved' pair, in line with studies on Western societies showing that "markedness is associated with evaluative negativity" (p.166). This asymmetry in (3) has the following implication in (4): using 'not saved' signals nothing special *because* using 'die' would have signaled something unusual, and conversely, using 'not die' signals something specific *because* using 'saved' would have signaled nothing special.

The relation of markedness can be highly situation-dependent, and expectations and goals are crucial factors of this situation-dependency (Schul 2011, p.170). To illustrate in the Asian Disease, the expectations and goals set up in the decision scenario implies a conflict in terms of markedness in the description of the consequences. Indeed, 600 people expected to be *killed* makes 'die' unmarked and 'saved' marked *contra* the general tendency in Western societies, while the goal to *save* people implies the reverse. If the expectation and/or the goal reverse, i.e., people are expected to survive and/or the goal is to kill them, the previous conflict in markedness

could either resolve (both expectations and goal in line with the general tendency) or its sources would reverse (expectation in line with the tendency but goals against it). Hence markedness relations seems difficult to be used for fine grained analyses of the minimal descriptive structure of a consequence, notably because the communication of social norms through linguistic terms is highly situation-dependent, i.e., is not intrinsic to the descriptions of the consequences.

Notice also the following asymmetry in the uses of negations within the Asian Disease. While 'not' negates the attribute of what is quantified in the sure program, another form of negation negates what is being quantified in the descriptions of the two consequences within all the versions of the Asian Disease (except the mixed one of Mellers and Locke), i.e., '*no* people will be saved' and '*no*body will die'. Given the tricky role of negation in human communication (see the next subsection's references), removing this asymmetry, e.g., by using '600 people will not be saved' and '600 people will not die', could yield surprising experimental results.

Another form of polarity has been argued to be pervasive in the uses of numerical expression because of the "scalar properties of the number system", that is, as Karl Halvor Teigen (2011) puts it, "smaller values [e.g., 200 people saved] are included in larger ones [e.g., 600 people saved], but not vice versa" (p.212 and p.202). This property has been shown to have side effects on the way people make predictions or evaluate the predictions made by others (see Teigen 2011). Predictions said to be 'exact' are, on the one hand, not often made by decision makers, and, on the other hand, when made by the decision modeler, more often evaluated as lower-bounded (e.g., 'at least...') than as upper-bounded (e.g., 'at most...') by decision makers. Notice that this should not have implications in the Asian Disease because the decision scenario asks subjects to "[a]ssume that the *exact* scientific estimate of the consequences of the programs are as follows [...]" (Tversky and Kahneman 1981, p.453, my emphasis). Teigen (2011) argues from experiments on monetary versions of the Asian Disease that this assumption does not hold.

Though Teigen's (2011) subjects had to estimate predictions, not to make choices, David Mandel (2014) used an equivalent to the Asian Disease strict framing problems with lives where subjects both estimated predictions and made choices. He found that when the exactness is left implicit in the consequences (i.e., explicited only in the decision scenario), subjects indeed tend to interpret the consequences as lower bounded, and hence, he argues, rationally reverse their preferences across frames; when this lower boundedness is made explicit in the descriptions of

the consequences ('at least 200 people will be saved', 'at least 400 people will die'), which are thus different descriptions of *different* consequences, the original framing effect is maintained; but it disappears when the exactness is made explicit in the descriptions of the consequences (e.g., 'exactly 200 people will be saved'). Contradicting this latter result, Christina Chick, Valerie Reyna and Jonathan Corbin (2015) found that the original framing effects remained by using the exact same procedure (i.e., 'exactly' in the description of the consequences) along with others (in other experiments) ensuring that subjects were interpreting all numerical information as exact (even if not explicitly stated in the descriptions of the consequences). A charitable interpretation of these contradictory results could be made by taking the experimenters' goals into account. Indeed, Mandel's goal was to argue against irrationality claims made about the effects in strict framing experiments, while Chick et al.'s was to maintain such a claim against Mandel's criticism. But this charitable interpretation is likely to be controversial. Indeed, it is akin to an experimenter effect (with which *both* parties would disagree), pointing to a very high level of situation-dependency in the leaking of goal discussed so far.[26]

Summing up this subsection, fined grained (i.e., high resolution) analyses of the minimal descriptive structure of consequences point to an extreme form of situation-dependency regarding their implications for framing *effects*. There is a pervasive polarity underlying each of the independently moving parts with interactions between themselves and with the goals, expectations and the like from the decision scenario. It can be argued that the emphasis on positive/negative polarities, along with the signal of a norm and deviations from it, are both conceptually well in line with the theoretical structure of prospect theory. An in-depth analysis of the communicative structure of a given decision situation seems necessary to pin-down the direction of polarity at play in a given choice. The latter cannot be simply read off from the description of a given consequence. The psychologists discussed in this subsection (and to a lesser extent in the previous one) provide conceptual and methodological tools for an eventual formal modeling in economics.[27]

[26]The controversy started from a failure to replicate Mandel's result by psychologists Leif Nelson, Joe Simmons and Uri Simonsohn (see the article on their scientific blog http://datacolada.org/2013/12/19/11-exactly-the-most-famous-framing-effect-is-robust-to-precise-wording/ and Mandel's response on his own blog https://sites.google.com/site/themandelian/data-colada; both last consulted 13/02/2016).

[27]Two remarks are worth making. Firstly, we have not discussed the role of (2) (i.e., what is being quantified, e.g., "people") in the minimal descriptive structure because, on the one hand, psychologists have not investigated it in the sense of, e.g., replacing 'people' by 'corpses', and on the other hand (and as already discussed), the structure of strict framing effects is robust to whether lives or money is at stake. Secondly, since we focus on

The next subsection discusses the theoretical background from linguistic and the philosophy of language that is shared by the psychologists discussed in this subsection in order to put their contributions in perspective with the communicative structure of choices developed in this dissertation.

4.3.3 Conversational *versus* communicative structure of choices

With the exception of Chick et al. (2015), one common characteristic of the psychologists discussed in the previous subsection is their (weak or strong) inspirations from the so-called logic of conversation formalized by philosopher Paul Grice, along with its developments in theoretical linguistics within the subfield of pragmatics by the neo-Griceans, especially Laurence Horn. The goal of discussing the shared background from linguistics and philosophy of language of these psychologists is twofold. On the one hand, it allows to illustrate how the *standard* interpretation of the Asian Disease depends not only on a relation of logical equivalence but also on a relation of conversational equivalence. On the other, it also allows to make more precise the implications of the theory of speech act from Searle that was constitutive of the development of the communicative structure of choices in this dissertation (cf. chapter 1).

Originally, Grice's (1975) logic of conversation is formalized by one principle and a set of maxims and supermaxims. The principle is one stating the cooperative nature of daily conversation, i.e., the existence of a goal shared by the participants to a given conversation, which is not necessarily fixed and may be only about the direction of the linguistic exchange. If the cooperative principle is accepted by the participants of a conversation, then a set of maxims and supermaxims can be used, by the speaker to mean more than he says (or writes) – performing *conversational implicatures* –, and by the hearer to infer information beyond what is said (or written), i.e., making *conversational inferences* about what is meant. Table 4.1 summarizes Grice's system (1975, pp.45-6).

The implications of this system for the analysis of the Asian Disease can be illustrated by focusing on the save frame of the original problem. When presenting the problem, the decision

external frames and the decision modeler, we kept discussions of the decision maker's internal frame at minimum. It shall nevertheless be remarked that a consensus seems to emerge on the main cognitive processes explaining framing effects from a purely internal frame perspective being the ones underlying the decision maker's *attention* (see e.g., Keren 2012; Moxey 2011; Teigen 2011).

Cooperative principle			
"Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged"			
Regarding "the quantity of information to be provided", the Maxims of Quantity are: "1. Make your contribution as informative as is required [...]. 2. Do not make your contribution more informative than is required."	Under the supermaxim "Try to make your contribution one that is true", the Maxims of Quality are: "1. Do not say what you believe to be false. 2. Do not say that for which you lack adequate evidence."	Regarding what is said but *not* how it is said, "Be relevant" is the single Maxim of Relation	Under the supermaxim "Be perspicuous" regarding *not* what is said but how it is said, the Maxims of manner are: "1. Avoid obscurity of expression. 2. Avoid ambiguity. 3. Be brief [...]. 4. Be orderly."

Table 4.1: Grice's (1975) system

modeler say, among other things, 'If program A is adopted, 200 people will be saved'; suppose that the decision maker chooses the sure program A by saying 'I favor program A'. By the maxim of relation "Be relevant" the decision modeler implicates that the 200 people that will be saved *are part of the 600 that are expected to be killed*, the decision maker infers this and then implicates that he favors program A *over program B* (and plausibly further implicates that *he would prefer that A be implemented in a real situation or would chose it himself if he had to etc.*). Hence the standard interpretation of the Asian Disease presupposes that the decision maker and the decision modeler abide by the cooperative principle and use some supermaxims and maxims (at least the one of relevance) both to mean and infer more than what is said.[28]

Horn's version of neo-Griceanism in theoretical linguistic pragmatics refines of the maxims of quantity through the implicit pairs of polarized quantitative scales used in the making of *scalar* implicatures (see, e.g., Horn 2006). These scales are sets of terms ordered from the strongest to the weakest. For instance, $< all, most, many, some >$ and $< none, hardly, any, few >$ are respectively positive and negative scales of quantifiers (Israel 2006, p.703; see this reference for a thorough discussion of 'Horn scales' focused on the issue of polarity). The use of a given term to express a proposition, e.g., '*most* people will die', *logically entails* the truth of same proposition with a weaker term (from the same scale or the other) as a lower bound, e.g., '*at least some* [or *few*] people will die', and *conversationally implicates* the negation of the same proposition with

[28] Grice (1975) discusses at some length the conditions under which the cooperative principle would not be respected, or would be respected but implying inconsistencies between maxims and thus rational violations of some. He later extended his system in various ways, as did many followers in various disciplines (see Davis 2014; Korta and Perry 2015). The work of Horn discusses here is one instance of these extensions.

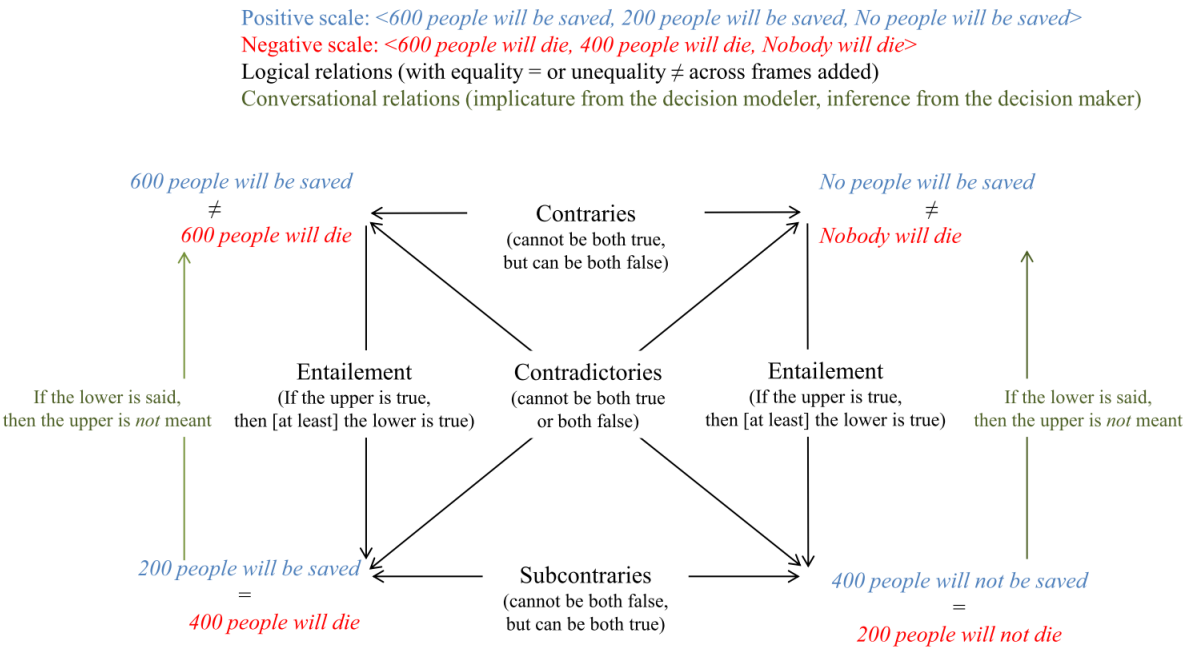

Figure 4.4: Post-Aristotlean's square of logical and conversational relations in the Asian Disease

a stronger term, e.g., '*not all* people will die'. Logical relations are deduced from the meaning of the terms while conversational ones such as scalar implicatures and inferences are induced from observed regularities in language uses. The conversational implicature just illustrated, i.e., using a weaker term to implicate (or infer) the negation of a stronger one is maybe the most robust existing regularities on Horn scales. Following Horn, we can use the so-called post-Aristotlean's square of opposition to illustrate the systematic contrast between logical and conversational relations among the terms of a given scale and their negations, taking the three consequences of the Asian Disease as terms of two ordered scales, i.e., one for the save frame and one for the die frame. In Figure 4.4, the vertical axis represents distinctions in quantitiy (universals up and particulars down), the horizontal one represents distinctions in quality (affirmations left and negations right), and relations are supposed to hold only for a given scale, i.e., only between pairs of same colored terms.

Notice again that the standard interpretation of the Asian Disease relies on a conversational implicature: the presentation of the consequence in the sure program *conversationally implicates* (rather than logically implies) the negation of the consequences in the probabilistic program.

Seeing the Asian Disease in this schema makes it clear that the interpretation of framing effects as irrational can be supported by pointing at *two* types of violations, namely of conversational and/or logical relations. Violations of conversational relations here are simply the non-inference of the conversational implicature just illustrated (in green on the schema) while violations of logical relations would her be a so-called fallacy of affirming the consequent, i.e., inferring from the truth of the entailed proposition that the entailing proposition is true. However, we have seen that the high degree of situation-dependency discussed in the two previous subsections allows several reasons to counter such irrationality claims, notably by taking into account the descriptive structure of the decision scenario and the decision question. Further reasons are provided by the even higher situation-dependency discussed by linguists on pragmatics of polarity and negation (e.g., Horn 2006; Israel 2006). This underlies rather subtle linguistic issues that we will not discuss here because it would carry the analysis too far away from the scope of decision theory, let alone economics.[29]

However the *conversational* perspective presented here can be contrasted with the *communicative* structure of choices underlying much arguments of this dissertation for the purpose of further developing the latter. To do so, we briefly discuss one incompatibility of the former with the theory of speech acts from Searle used in chapter 1 to flesh out the latter. This incompatibility concerns the notion of meaning underlying Grice's logic of conversation (Grice 1957; Searle 1965; 1969, §2.6; 2007). Grice's notion of meaning is intentional and self-referential in the following sense. A speaker has meant something to a hearer by saying something else *if* the former had the intention to produce an effect (e.g., a belief, desire etc.) on the latter *by* getting him to recognize that very intention to produce that very effect. To illustrate with the Asian Disease, a decision modeler meaning 'choose between the certain and the probabilistic programs' consists in producing the belief (i.e., the effect) 'I have to choose between the two

[29]Rubinstein has argued that the main obstacle for such potential relevance is the necessity of the cooperative principle, which severely limits the set of human interactions to be studied (see Rubinstein 2000, chap.3; 2012, chap.4). To bypass this obstacle, Rubinstein uses game theory to model linguistic exchanges of arguments in persuasion situations. It is not straightforward to apply his analysis to framing effects because, as a byproduct of game theory, there is an assumption of common knowledge about the decision situation: "[w]hen the listener interprets a statement in a persuasion situation, he is aware of the fact that the persuader is trying hard to convince him" (2012, p.192). That is one reason why a decision theoretic approach to the problem remains relevant, despite the strong game theoretic taste of the considerations discussed so far. See del Corral and Bonilla (2008) for further references on pragmatics and game theory. Conversational pragmatics could also be used in economics could because the underlying logic and mathematics of Horn scales are very close to the mathematics from measurement theory used in decision theory (see Geurts 2013 for a contribution by a philosopher in that direction). However, the extent to which Horn scales can be used without the cooperative principle is unclear.

programs' by getting the decision maker to recognize his intention to create that very belief *by* saying 'which of the two program would you favor?'. Though Searle uses Grice's notion of meaning in his theory of speech act, he revises it on two points he thinks are defective.

Firstly, meaning something is not always and only a matter of *intention*, it can also be a matter of *convention*, i.e., what the words mean in a given language independent of the speaker's intention (Searle 1969, pp.44-45). For instance, imagine that I am a decision modeler who prefers Program B but has to recommend Program A to a decision maker (e.g., as in van Buiten and Keren 2009). To do so, I intend to produce both the belief that I prefer Program A and the desire to choose it by describing it as '200 people will be saved'. Suppose further that the decision maker recognizes my intention and this creates the aforementioned belief and desire, leading to his choice of Program A. Does it follow that I *meant* 'I prefer Program A' by saying '200 people will be saved'? No, what I meant was the conventional meaning of these words, namely that 200 people will be saved (and 400 will die out of the 600). Notice that the principle of cooperation is violated in this situation. Hence, for the purpose of making value judgments of rationality or irrationality about a given framing effect, an important feature to look for in the communicative structure of choices is the asymmetry created by the decision modeler's unilateral violations of the cooperative principle through a calculated balance between intention and convention in the performance of his illocutionary acts.

Secondly, Grice's emphasis on the effects produced on the hearer blurs a crucial distinction proposed by Austin (1975) in the theory of speech act, between so-called *illocutionary* act, i.e., the type of speech act performed, and *perlocutionary* act, i.e., the effects produced on the hearer by the performance of a speech act (Searle 1969, pp.46-47). Types of illocutionary acts include: making a statement or a promise or an order etc. Types of perlocutionary acts include: being convinced (of a statement) or being expectant (of a promise) or being subordinate (of an order). Searle argues that what a speaker means is what he does in the performance of the illocutionary act, but not necessarily what happens through the perlocutionary act. For instance, a decision modeler can mean that '200 people will be saved' by stating it without convincing an excessively stubborn decision maker of the truth of this statement (just as an excessively skeptic or unsubmissive hearer may never become expectant or subordinate, respectively). Given that we are interested primarily by the decision maker's behavior, it should therefore be noted that

the primitive empirical phenomena we are looking for in the communicative structure of choices are different perlocutionary acts achieved by different descriptions of a same consequence. As remarked by several authors, perlocution is by far the most underdeveloped notion in the theory of speech act, notably due to Searle's focus on illocution (see Davis 1980; Gu 1993; Attardo 1997; Kurzon 1998; Marcu 2000, Kang 2013; and esp. Liu 2013). One can speculate that some fruitful mutual developments are possible between speech act theory and decision theory through the study of (at least 'strict') framing effects.[30]

Summing up this subsection, the Gricean background underlying the contributions from the previous subsection is not entirely compatible with the communicative structure of choices as developed in this dissertation, notably through Searle's theory of speech act. Nevertheless, it is useful to points how the standard interpretation of the Asian Disease does not rely only logical relations but also on conversational ones *and* to flesh out further implications of our communicative structure of choices.

Conclusion

At a very general level, the main argument of this section can be put in an information theoretic fashion. The decision modeler's *choice* of one possible external frame over another may send signals to the decision maker in the form of tacit information about the decision situation that are not contained in the literal messages constituting the decision problem. The effects of this tacit informational surplus cannot be formally captured in standard models without relaxing the axiom of description invariance. The latter can be made formally explicit at least at two different levels of resolution. One level higher than the standard model would capture different descriptions of the same consequence. One level even higher would capture different combinations of the independently moving parts constitutive of the minimal descriptive structure of a consequence. However, the subtleties of ordinary language uses at the latter level seem too complex to be tractable. From the perspective of axiomatic decision theory, it seems wise to stay at the former level just to capture that different descriptions of a consequence can induce different preferences (rationally or not). From the perspective of modelization, the latter level may be

[30] On Austin's account every speech act is also constituted by the *locutionary* act of pronouncing sounds or writing symbols according to syntactical and semantical rules.

appropriate to capture the complex situation-dependency through a perlocution parameter.

Conclusion and transition: on the meaning of a 'consequence' in economics

This chapter proposed a systematic study of framing phenomena focused on strict framings of consequences. Regardless of the normative implications, preference reversals triggered by different descriptions of the same consequences is a pervasive and robust phenomenon. The pervasiveness of violations of description invariance is already well admitted around behavioral economics. That these violations are not necessarily irrational is not as well recognized. Two conditions under which such violations can be seen as rational are because the experienced utility of a consequence may depend on its description and/or the decision modeler's choice of one description over another may itself be a choice-relevant information for the decision maker. Fully fleshing out the normative implications of the latter is a complex task due to the subtle linguistic details of the communicative structure of choices. How are these conclusions extendable to further framing phenomena? If framing phenomena is broadly defined as preference reversals triggered by different presentations of the same decision problem, then much of the conclusions may extend quite straightforwardly. It does not seem totally off the hook to conceive systematic dependence of experienced utility on the presentation of a decision problem and/or systematic leakage of choice relevant information by the decision modeler through his choice of one presentation over another. However, much of the other framing phenomena are not strict but loose, i.e., they involve pairs of different but equivalent decision problems, not pairs of different descriptions of the same decision problem. Because loose framing effects still involve ordinary language uses to make different *descriptions* (of features of *different* problems), the linguistic details of the communicative structure of choices should still matter. There is no reason to believe that the pervasive situation-dependency highlighted here for strict framing effects should not carry over to loose framing effects. But because loose framings involve extra-linguistic differences across frames, the normative implications of preference reversals may be easier to draw than with purely linguistic differences, as illustrated with the framings under certainty and of acts and contingencies under uncertainty.

At several points throughout the chapter, a parallel was made between, on the one hand, the Asian Disease and the axiom of description invariance, and, on the other hand, the Allais paradoxes and the axiom of independence. Notably, both were initially implicit in the standard model and the importance of making them explicit was mostly motivated by experimental demonstrations of their violations. However, the independence axiom greatly outnumber the axiom of description invariance in terms of existing publications that (1) make the axiom formally explicit, (2) weaken it to account for its violation and (3) defend such weakening as rational. The intended contribution of the next chapter will be to use the systematic study in this one to do those three things for the axiom of description invariance. By way of transition, I would like to discuss a last parallel with the independence axiom. Looking back, some of the controversies around the Allais paradoxes revolved around the descriptive structure of a consequence. As reviewed by Broome (1991, chap.5), most defenses of expected utility theory consisted in arguing that Allais' paradoxes are not paradoxes anymore whence the consequences are *redescribed* to add information about, for instance, some feelings of disappointment if the decision maker does not win anything. Such redescriptions, e.g., '0' and '0 plus disappointment' are then taken as *different* descriptions of *different* consequences, hence breaking the equivalence between the two initial pairs of problem, and making it impossible to observe an inconsistency from expected utility theory. Shouldn't the same reasoning apply to the discussion of the Asian Disease conducted in this chapter? It could, but a case can be made against it.[31]

The source of this reasoning (i.e., redescription with subjective elements) is the standard notion of a consequence, which, following Savage (1972 [1954], p.13), is taken to mean "anything that may happen to the person", i.e., to the decision maker. This includes any "things, or experiences, regarded as consequences" (ibid, p.14). It may be asked, regarded as a consequence by whom? In standard decision theory, consequences are, *in principle*, regarded as such by decision makers. But by virtue of revealed preference methodology, only the decision maker's observable choice behavior should be a primitive in the analysis which excludes his subjective perception of what counts as one consequences. Hence, *in practice*, consequences are regarded

[31] It should be noted that Savage's (1972 [1954], p.103) famous response to Allais does not involve redescriptions of the consequences, but additions of events to 'materialize' Allais' probabilities (from abstract numbers to lottery ticket numbers). Savage then argued that his version and Allais' original one were two different descriptions of *the same* decision problem, and that his preferences in his version made him change his original preferences in Allais' original version.

as such by a decision theorist guessing or postulating what a decision maker would count as a consequence. In a way, the rather permissive and subjective meaning of a consequence is a necessary requirement for the shared *p*sychology between the economist and the economic agent to be effective in the practice of decision theory. But the decision theorist's eventual redescription of the Asian Disease's consequences would prevents to account for the simple fact that *one* decision maker (or even *one* decision theorist for that matter) can entertain *different* perspectives on the same consequence. That is, it prevents to account for the recognition by the decision maker that different presentations of the same consequence (or "things") trigger different preferences (or "experiences"). Instead, it can be argued that we shall try to do justice to the subjectivity of a decision maker who recognizes that one and the same consequence can trigger different mental states in different situation, without conveniently postulating that the differences were 'in fact' (or 'equivalently') about the consequences. This is in line with Sen's position (notably within the debates around the independence axiom) not to blend all the factually relevant observations into one single theoretical construct (here of 'consequence') in order to do justice to the decision maker's values and hence make the related normative analysis easier (see esp. Sen 1979; 1980; 1982; 2002, chaps.6, 11 and 15). In short, that different descriptions of one same consequence can be described as such by a decision modeler and recognized as such by a decision maker should be represented as such by the decision theorist. The next chapter proposes one way of doing that.

Chapter 5

An axiomatic framework to deal with framing effects[1]

In this chapter, our goal is to provide an axiomatic structure in which some of the theoretical implications of framing effects discussed in the previous chapter can be formally dealt with. This is tantamount to making the axiom of description invariance formally explicit and then to weaken it in a way that allows the representation of framing effects in terms of choice, preference, utility and the equivalence among them. Our intended contribution is to show that description invariance is not necessarily an 'all or nothing' axiom: either you have it or you don't. It can be characterized so that only some of its parts can be dropped, without dropping the whole axiom. In other words, our goal is to characterize the role of language in economic rationality through the technical language of economics, which implies using formal language to allow standard and behavioral models to account for the uses of ordinary language by economic agents.

To do so, we broaden a little bit the interpretation of the decision modeler, here denoted by d_{mo}, who is not necessarily the one who poses a decision problem, but the one whose perspective on a decision problem *may* be different from the perspective of the decision maker, here denoted by d_{ma}. Hence, in the applications of our framework, the d_{mo} *can* be identified with the one

[1] Most of this chapter is taken from a working paper co-written with Dino Borie who practices axiomatic decision theory. In line with the literature on decision theory in economics, the word 'axiom' is used in its technical sense of being necessarily true *for a given deduction* (of a proposition, theorem etc.), not in its old technical sense of being necessarily true *in the absolute* by contrast with a 'postulate' which is true for a given purpose or in a given science; in modern axiomatic work by logicians, mathematicians and scientists, the distinction between axiom and postulate seems to be no more in use, both are called axiom (see Mongin 2003, esp. p.107). Hence it make sense to say that an axiom holds for a given deduction but needs to be weakened for another one.

who poses the decision problem (an experimenter, a policy maker, a firm, one particular self of a multiple selves d_{ma} etc.) but it can also be identified with the one who just observes the decision situation (an economist, a psychologist, a politician etc.). This broadening captures the main condition of possibility for the existence of framing effects, namely that there can be *different* perspectives on the *same* decision situation.[2]

Our axiomatic framework essentially refines the formal structure of the objects of choice within the standard model so as to account for the specificity of cases where they are partly constituted of ordinary language uses. In a nutshell, we construct a choice set with descriptions of objects of choice instead of objects of choice, an equivalence relation determining which descriptions are about the same object of choice from the d_{mo}'s perspective and a concatenation operation joining different equivalent descriptions to form a new description (e.g., "200 people will be saved" can be concatenated with "400 people will not be saved" to form "200 people will be saved and thus 400 people will not be saved"). The d_{ma}'s preferences over descriptions of objects of choice will be a standard weak order. Our results determine how the d_{mo}'s equivalence relation is articulated with the d_{ma}'s indifference relation. This allows to characterize the key implicit axiom of the standard model, which we call "Independence of Common Description" (ICD), that makes it blind to framing effects we are interested in and to the communicative structure of choices. We weaken the latter to allow such effects to be formally represented. Further axioms characterizing finer degrees of dependence to descriptions are distinguished, along with informational measures derivable from our framework.

Recall the fivefold picture presented in the previous chapter to characterize existing discussions of description invariance in economics: (1) description invariance is an implicit axiom of the standard model, the violations of which are (2) descriptively pervasive, (3) normatively unjustifiable, (4) mathematically intractable and (5) explained by prospect theory. Our intended contribution can be summarized with respect to these five points. (1) Description invariance is an implicit axiom of the standard model that we make explicit by working primarily on the formal structure of its object of choice and only derivatively on the d_{ma}'s preferences. We are

[2] From a methodological perspective, this broader interpretation of the relation between the decision modeler and the decision maker who may or may not share a common ordinary language and *p*sychology seems similar to the position defended by Fritz Machlup (1978, chap.12). Likewise, the emphasis on a systematic distinction between the economist and the economic agent to deal with issues underlying individual rationality in economics seems similar to the position defended by Joseph Schumpeter (see Festré and Garrouste 2008; Arena 2012). A more recent account of somewhat similar position is defended by Larry Samuelson (2005).

driven by the conditions under which it is violated in psychologists' experiments on the Asian Disease (2) to make it formally explicit, (3) to provide normative justifications for weakening it, and (4) to do so in a mathematically tractable way. (5) We account for variations in framing effects for which prospect theory cannot (at least straightforwardly) account for.

More precisely, we tackle the issue of interdisciplinarity by intending our framework to contribute to economics while motivating our axiomatic constructions with evidence from *P*sychology. We focus on three subsets of evidence from all the ones that were discussed in the previous chapter. The first one includes the effects of further redescriptions of the same consequence, i.e., the last set of results presented in §4.2.2: "200 people will be saved and 400 people will not be saved", "400 people will not be saved", etc. The second one includes the direct impact of the description of an object of choice on the utility derived from the occurrence of its associated consequence, e.g., being better off when the rescue of 200 people and the death of 400 occurs from having chosen A than from having chosen C: see the penultimate results presented in §4.2.3. The last one includes the results presented in §4.3.1 about how the decision modeler makes choices from sets of different descriptions of the same object of choice and how these very choices (of one description over other possible ones) may influence the decision maker, e.g., the former may signal that he would prefer that the sure program be chosen by describing it as "200 people will be saved" instead of "400 people will die" and the decision maker may infer that signal.

We tackle the positive/normative issue by showing how the framework allows, in applications, for clarifications and more precise discussions of the conditions under which a framing effects is rational or irrational. By virtue of the distinction between the d_{mo}'s and d_{ma}'s perspectives, our framework is not tied to a universal position whereby they are all necessarily either rational or irrational. We emphasize in the conclusion that though our uses of the terms 'rational' and 'irrational' are not defined 'once and for all' *a priori*, their meanings are justified situations by situations.

This chapter is structured in three sections. We first present the axioms and results characterizing our axiomatic framework (5.1). We then discuss the representations terms of utility and choice functions possible from it (5.2). And we conduct a general discussion with several applications, possible extensions and implications for the positive/normative issue underlying

framing effects (5.3).[3]

5.1 Axioms and results

We first present the primitives of the framework (5.1.1). We then provide an axiomatic characterization of the standard models' unframed descriptive structure and suggest how to weaken it to allow for a formal representation of framing effects compatible with empirical results and intuitions. To do so, we present the basic axioms that we never weaken (5.1.2) before the ones that characterize description invariance (5.1.3) which we weaken (5.1.4) in a way that allow us to characterize two degrees of description dependence (5.1.5).

5.1.1 Primitives

Our framework has four primitives. The first one is X the nonempty set of objects of choice. Its elements are descriptions of objects of choice instead of objects of choice. For the construction of the framework, our objects of choice are going to be "described consequences", in contrast to the standard model's "consequences". This is the first step need to characterize description invariance in terms of an equivalence between our framework and the standard model. Furthermore, we will work on static decision problems under certainty because violations of description invariance can occur at this basic level, e.g., the attribute framing discussed above and other framing effects discussed later. Abstracting from risk, the Asian Disease provides a handy illustrative example throughout, which we will use with the following notation. Consequences

[3]It should be noted that we stick to the term 'description invariance' as in the previous chapter and as originally used by Kahneman and Tversky (1984), while some other authors uses 'extensionality' and even sometimes 'consequentialism' (e.g., Tversky and Kahneman 1986, S253) to refer to what framing effects violated in standard models. Though it is beyond the scope of our framework to analyze how exactly are description invariance, extensionality and consequentialism related within standard models, here are a few hints. It can be argued that the 'invariance' in 'description invariance' refers to the mathematics of measurement theory, in the making of which Tversky played a certain role (see Heukelom 2014). In this context, invariance roughly means: what needs to remain invariant in a measurement procedure for the numerical outcome to be meaningful (see Luce et al., 1990, chapter 22; Chao, 2007; Boumans, 2005, 144-119; 2007). By contrast, 'extensionality' was first used in a discussion of framing by Arrow (1982), whose well-known background in the symbolic logic of Tarsky easily led us to infer that extensionality here indeed refers to a famous foundational issue in the philosophy of logic. Roughly, the issue is about the establishment of equivalence relations (and about the thorny issue of defining 'logical consequences'), and seem to have always been controversial (see e.g., Chisholm, 1941, sect. III; Anscombe, 1969; Quine, 1994; Thiel, 2003, 2009). And so have been the implications of 'extensionality' for economics (see e.g., Bacharach, 1994; Vilks, 1995; Cubitt and Sugden, 2003; Moscati, 2012). And so have also been its implications in psychology within the debates on the psychology of inductive and deductive reasoning in which Kahneman and Tversky's work are central (see e.g., Cohen, 1981). Finally, Kahneman and Tversky themselves refer to the work of Hammond (1988a; b) on consequentialism, whose role in decision theory we have briefly presented in chapter 2.

Consequences	Described consequences
$[x]$	$x_1 \equiv$ "200 people will be saved" $x_2 \equiv$ "400 people will die" $x_3 \equiv$ "400 people will not be saved" $x_4 \equiv$ "200 people will not die" $\vdots$
$[y]$	$y_1 \equiv$ "600 people will be saved" $y_2 \equiv$ "nobody will die" $y_3 \equiv$ "600 people will not be saved" $y_4 \equiv$ "600 people will not die" $\vdots$
$[z]$	$z_1 \equiv$ "no people will be saved" $z_2 \equiv$ "600 people will die" $z_3 \equiv$ "600 will not be saved" $z_4 \equiv$ "600 people will not die" $\vdots$

Table 5.1: Consequences and described consequences in the Asian Disease

are denoted by letters within brackets, e.g., $[x]$, described consequences are denoted by letters with subscripts differentiating different descriptions of a same consequence, e.g., x_1, x_2, and the symbol $\equiv$ reads 'can be empirically instantiated as'. Table 5.1 shows how to use these symbols with respect to most of the descriptions discussed in the previous chapter.

The second primitive is a nontrivial binary relation $\approx$ on X. Two described consequences related by this binary relation are such that, from the "decision modeler" (d_{mo}) perspective, they are two descriptions of the same consequence (and conversely). Observe that $\approx$, by definition, is an equivalence relation; if $x \approx y$, we say that x and y are *consequentially equivalent*. In other words, the d_{mo}'s perspective defines equivalence classes of described consequences. By contrast with the standard model where it is implicitly assumed that d_{ma} and d_{mo} share the same perspective on what counts as a consequence, we want the d_{ma}'s perspective to be analytically free from the d_{mo}'s, at least within the primitives of our model.

The third primitive is a partial binary operation $\circ$ on X. If $x \approx y$ we may wish to consider them, taken together, as forming a single described consequence (i.e., a single object of choice under certainty) denoted $x \circ y$ and called the *concatenation* of x and y. The concatenation $x \circ y$

is the joint description of the described consequences x and y, and is consequentially equivalent to either one of them. To illustrate with the Asian Disease example, if the d_{mo} considers that "200 people will be saved" (x_1) and "400 people will not be saved" (x_3) are two different descriptions of the same consequences, i.e., are consequentially equivalent ($x_1 \approx x_3$), then they can be concatenated and the concatenation is itself consequentially equivalent to either one of the two described consequences. In other words, we have three different descriptions of the same consequence: x_1, x_3 and $x_1 \circ x_3 \equiv$ "200 people will be saved and 400 people will not be saved".

Definition 1. Let X be a nonempty set, $\approx$ a nontrivial binary relation on X, and $\circ$ a partial binary operation on X. The triple $\langle X, \approx, \circ \rangle$ is a *descriptive structure* iff $\approx$ is an equivalence relation and for all $x, y \in X$, if $x \approx y$, then $x \circ y$ is defined, and $x \circ y \approx x$.

Three related points need to be emphasized. First, the concatenation of described consequences is different from the concatenation of consequences. The former is akin to the combination of information about one consequence. The latter is akin to the combination of at least two consequences. This contrast can be illustrated by a comparison with the operation of joint receipt (Luce and Fishburn 1991). While the joint description of one glass of water "half-full" with one glass of water "half-empty" is the description of a glass of water "half-full half-empty", the joint receipt of one glass of water half-full with one glass of water half-empty is a full glass of water. [4]

Second, the operation of concatenation is not to be confused with the operations of addition or multiplication (it can only define these operations if more structure is added). Thus, concatenation of descriptions *can* imply the addition of information from the d_{ma}'s perspective but not from the d_{mo}'s; concateation just explicitly puts next to each other what the d_{mo} takes to be consequentially equivalent. In other words, the concatenation of information *can* imply an asymmetry of information between the d_{ma} and the d_{mo}.

Third, the operation of concatenation does not imply a one-to-one translation between (1) the mathematical representations of described consequences in the formal language of our setup and (2) the empirical description of consequences in the ordinary language of, say, an experimental

[4] Assuming fungibility of water, which seems less problematic than fungibility of money, and that consequences are only about water, not its physical container.

setup testing framing effects. This is not the case because, by contrast with formal languages, the concatenation of two described consequences in ordinary language is more vague (there are multiple representations of a concatenation) but also more efficient (recursivity within the representations avoids redundant repetitions which decreases the amount of symbols). For instance, take $x \equiv$ "a glass of water half-full" and $y \equiv$ "a glass of water half-empty". In ordinary language, it would be natural that $x \circ y \equiv$ "a glass of water half-full and half-empty" or "a glass of water half-full or half-empty" or "an half-full half-empty glass of water " or yet other formulations. But it is very unlikely that such formulations include "a glass of water half-full a glass of water half-empty". These are confusing, and implied by the requirement of a one-to-one translation between our framework and its empirical counterpart. Therefore, our framework is agnostic about the *exact* empirical counterparts of the concatenation operation $\circ$. This is in part why we said that $\equiv$ reads as '*can* be empirically instantiated as' (and not 'is necessarily instantiated as'), which allows $\circ$ to be empirically described by "and", "or", ",", " " (just a space), etc. We just want to formally capture the possibility condition of framing effects: the simple fact that one consequence can be presented under different descriptions, and that some of these descriptions are combinations of other descriptions.

Nevertheless, common sense derived from our everyday uses of ordinary language, and, most importantly, the large experimental literature on framing effects, provide fairly clear guidance for the empirical application of our framework. As an illustration, consider the structure of the sure thing in Asian Disease experiments. A described consequence involves expressions of four dimensions of the consequence, i.e., (1) of a quantity ("200", "400"), (2) of what is being quantified ("people"), (3) of an attribute of what is being quantified ("saved", "die") and (4) of the implicit assertion of the attribute ($\emptyset$) or the explicit negation of it ("will $\emptyset$ be saved"/"will $\emptyset$ die", "will not be saved"/"will not die"). Empirically speaking, we have 'minimal descriptive structure' of a consequence under certainty from the previous chapter, which requires a combination of at least one expression of each dimension. Concatenations of described consequences often involve the repetition, and always requires a combination, of at least two different expressions of (at least) one dimension: "200 people will be saved and 400 people will not be saved", (or possibly) "400 people will die and 200 will not". Where we are formally agnostic is on the empirical manifestation of those dimensions, the repetition of those

dimensions, and the combination order of those dimensions.

To be clear on the interpretations of our primitives, let us contrast three possibilities. Firstly, if X would be the set of all typographical symbols, the concatenation operation could be interpreted as forming words and sentences, e.g., $w \circ i \circ l \circ l \circ \circ d \circ i \circ e \equiv$ "will die". Secondly, if X would be the set of all expressions of the four dimensions of the minimal descriptive structure of a consequence, the concatenation operation could be interpreted as forming described consequences, e.g., with $x \equiv$ "200", $y \equiv$ "people" $z \equiv$ "will be saved", $x \circ y \circ z \equiv$ "200 people will be saved". Clearly, both these levels of analysis are not fitted for economics. Hence, the third interpretation is the one we take in this chapter, where X is the set of described consequences, empirically corresponding the expressions of *at least* one instance of each of the four dimensions of the minimal descriptive structure of a consequence.[5]

What we want to formalize is that described consequences can *generate* further described consequences, by concatenation. Formally, let $\langle X, \approx, \circ \rangle$ be a descriptive structure, $x \in X$ be a described consequence and $[x]_\approx := \{y \in X \mid y \approx x\}$ be the equivalence class of x by $\approx$.

Definition 2. $[x]_\approx$ is said to be generated by $D \subset [x]_\approx$ iff every element of $[x]_\approx$ can be expressed as the concatenation of finitely many elements of D; and when both conditions hold, $[x]_\approx = \langle D \rangle$ and the elements in D are called *generators*.

Generators are simply described consequences that are not the result of a concatenation of, and thus cannot be broken into, other described consequences. Identifying potential generators in the experimental study of framing effects is useful to get a picture of the conditions under which the effects hold and vary across different descriptions of the same consequences. For example, in the literature on the Asian Disease problem, D contains at least four generators of the sure program, namely x_1, x_2, x_3 and x_4 in Table 5.1. Their concatenations generate further descriptions of the sure program, as shown in Table 5.2.

[5]Maybe the former level would be more fitted for linguists, logicians or philosopher of language. And the work done by psycholinguists or psychologists inspired by psycholinguistics on framing effects can be interpreted as been done at the latter level (Keren 2011b, esp. part II). Though some of their results may be useful for further modeling or empirical work, there is a potential lack of tractability worth noting due to, as *they* show, the potentially infinite makings of synonymous expressions (e.g., "alive", "perish") and the complex asymmetries between negations and contraries.

	Kahneman and Tversky (1981)	Kühberger (1995, exp.1)	Kühberger (1995, exp.2)	Mellers and Locke (2007, p.362)
As	x_1	x_3	$x_1 \circ x_3$	$x_1 \circ x_2$
Cs	x_2	x_4	$x_2 \circ x_4$	

Table 5.2: Generators and concatenations in the Asian Disease

The last primitive is the d_{ma}'s preference relation $\succsim$ on X. $x \succsim y$ means that x is at least as desirable as y. As usual, the strict preference $\succ$ and indifference $\sim$ relations are respectively defined as the asymmetric and symmetric parts of $\succsim$. We emphasize that $\succsim$ is defined over described consequences. This contrasts with the standard model, where d_{ma}'s preferences are defined over consequences. That is, in the standard model, the d_{ma}'s preferences are implicitly defined over the d_{mo}'s partition of X, i.e., on $X/\approx$, the set of equivalence classes of X under $\approx$. Notice that by restricting the concatenation of described consequences to the d_{mo}'s perspective (i.e., the operation of concatenation can only be performed between what is consequentially equivalent, that is, between what is contained within equivalence classes of X defined by $\approx$), such concatenations are logically independent from d_{ma}'s preferences, which are defined over X i.e., over all described consequences irrespective of their equivalence classes from the d_{mo}'s perspective. This can be seen as a prerequisite to avoid the standard model's problem of logical omniscience, i.e., that the d_{ma} knows every logical implications of his knowledge (Lipman 1999); or more precisely, that the d_{ma} knows and agrees with every implications of his knowledge derivable from the d_{mo}'s logic when the latter is identified with a decision theorist. That furthermore implies, in the standard model, that the d_{ma}'s indifference relation implicitly holds among all the elements of a given equivalence class, that is, among all different descriptions of the same consequence.[6]

Within the present framework, it is possible to characterize three relevant cases of preferences under framing with different degrees of discrimination of consequences. The first case

[6]Logical omniscience is however plausible in *some* situations such as the glass example. If a glass of water is described as half-full to the d_{ma}, it is very likely that he can unproblematically infer that the glass is also half-empty. But this is less plausible in other more complex and economically relevant situations such as the Asian Disease examples, and other ones discussed in the literature on framing effects. A non-trivial and intuitive example of such situations involve inflation, where the d_{mo}'s perspective establishes *equivalence* of purchasing powers (i.e., economic *consequences*) in terms of income and prices, but that equivalence is often not shared from d_{ma}'s perspective; i.e., the inference of equivalent purchasing powers at different levels of income and prices (the combinations of which are then just different descriptions of the same economic consequences) is not so straightforward. See Arrow (1982) for an early discussion, and Shafir et al. (1997) for empirical support and further discussion.

corresponds to what is usually called *description invariance* in the literature. This axiom is invisible in the standard model because the latter works in an *unframed descriptive structure* where preferences are only representable over consequences, e.g., over the d_{mo}'s equivalence classes of consequences $[x]_{\approx}$, $[y]_{\approx}$, etc. but not over described consequences, i.e., over x_1, y_1, x_2, y_2, etc. By virtue of the structure of our choice set, we are always working in a *framed descriptive structure* where preferences over described consequences are representable, hence description invariance can be made formally explicit. The standard model's case where it is impossible to discriminate different descriptions of a given consequence corresponds here to the case where the d_{ma}'s indifference relation holds among all the elements of a given equivalence class:[7]

$$\forall x, y \in X, \quad x \approx y \Rightarrow x \sim y.$$

The second case is when regardless of the d_{ma}'s preferences among all possible descriptions of the same consequence, his preferences among different consequences are "ordered". We call this case *tidy description-dependence*. Formally,

$$\forall x, y, z \in X, \ \textit{suppose } x \approx z \textit{ and } x \succsim z, \ \textit{then } x \succsim y \textit{ and } y \succsim z \Rightarrow y \approx x.$$

Remember that we are under certainty here, hence how the original framing effect in the Asian Disease can be captured by tidy dependence cannot be illustrated directly. To do that, we need to introduce probability distributions in our framework, which is done in the applications. Tidy description-dependence prevents cases where the d_{ma} prefers one description of a consequence to another description of *another* consequence and yet prefers a further description of the latter consequence to a further description of the former (i.e., a framing effect under certainty, e.g., from Table 5.1, $y_1 \succsim x_1$ and $x_4 \succsim y_2$). This kind of case is the third one we distinguish here and call *untidy description-dependence*. We characterize it only informally because there is no formal relation within the set of consequences and there is no characterizable logic without further hypotheses about the d_{ma}'s decision making process.

[7] We are not the first one to *state* description independence in this way, see, e.g., Bacharach (2003, p.65), Blume et al. (2013, p.9), Lerner (2014, p.40) and esp. Giraud (2004b, p.52). We discuss these in the section on applications, where the influence of Giraud's work on our framework is acknowledged. It can be argued that '*the*' standard model is in fact characterized by $x \approx y \Rightarrow [x] \sim [y]$. If we would have defined preferences over consequences, this would be the natural way of characterizing description invariance; however, with our preference defined over described consequences, we feel that the above characterization is more natural.

Hence it can be argued from a methodological perspective that, in our framework, the framed descriptive structure with description invariance always characterize standard economics and the normative positions of behavioral economics on framing effects. The framed descriptive structure with tidy or untidy description-dependence marks different ways of seeing the frontiers between standard and behavioral economics on framing effects. Under certainty, tidy descriptive dependence never yield preference reversals, hence it may be acceptable to some standard and behavioral economists as rational. Under risk, however, it can yield preference reversals, which are usually deemed as irrational by behavioral and standard economists, but needs to be captured in positive models of individual behaviors in behavioral economics. This reasoning always applies to untidy description-dependence, i.e., regardless of whether or not we are under certainty or under risk.

5.1.2 Basic axioms

We first state our axioms and then emphasize their empirical interpretations. We use a structure inspired by the so-called "extensive structure" from Krantz et al. (1971) – the distinctions being that we add the idempotency axiom, our concatenation operation is not total and we do not use any Archimedean axioms.

Axiom 1 (Descriptive structure). $\langle X, \approx, \circ \rangle$ *satisfies Definition 1.*

Axiom 2 (Weak order). $\succsim$ *is complete and transitive.*

Axiom 3 (Weak commutativity). *If* $x \circ y$ *and* $y \circ x$ *are defined, then* $x \circ y \sim y \circ x$.

Axiom 4 (Weak associativity). *If* $(x \circ y) \circ z$ *and* $x \circ (y \circ t)$ *are defined, then* $(x \circ y) \circ z \sim x \circ (y \circ z)$.

Axiom 5 (Weak idempotency). *If* $x \circ x$ *is defined, then* $x \circ x \sim x$.

Axiom 1 corresponds to the structure of objects of choice from the d_{mo}'s point of view. Its empirical interpretation has already been given above. Axiom 2 is a standard requirement

about d_{ma}'s preference relation over the objects of choice, that is, his preference relation is a weak order. Notice that description invariance is not implied by the weak order alone. It is implied by the behavior of its indifference part over possible combination of descriptions, as characterized by the other axioms. Axioms 3 and 4 state that whenever the concatenation is defined, it does not matter for the d_{ma} in what order described consequences are concatenated. The order of the described consequences being concatenated (weak commutativity) and the order of the concatenations of described consequences (weak associativity) may be reversed without altering informational content, hence the d_{ma}'s indifference. Notice that these axioms are plausible within the scope of the framing effects with which we are concerned; but would need to be relaxed to study order effects under framing, with which we are not concerned (in any case, empirically, it has been shown that there is no impact of the order of presentation in framing effects of the Asian Disease type).[8]

Example 1. From Table 5.1, $x_1 \circ x_3$, $x_3 \circ x_1$, $(x_1 \circ x_3) \circ x_2$ and $x_1 \circ (x_3 \circ x_2)$ can be defined by the d_{mo} to correspond to:

$x_1 \circ x_3 \equiv$ "200 people will be saved and 400 people will not be saved",

$x_3 \circ x_1 \equiv$ "400 people will not be saved and 200 people will be saved".

and

$(x_1 \circ x_3) \circ x_2 \equiv$ "200 people will be saved and 400 people will not be saved and also 400 people will die",

$x_1 \circ (x_3 \circ x_2) \equiv$ "200 people will be saved and also 400 people will not be saved and 400 people will die".

Hence, what remains implicit in the standard model is that because all the generators (x_1, x_2, x_3) and the concatenations are consequentially equivalent, the d_{ma} is indifferent between the order of the described consequences being concatenated ($x_1 \circ x_3 \sim x_3 \circ x_1$ by

[8]Notice that axiom 2 can be weaken to a preorder (reflexive and transitive binary relation) quite easily. In that case, axioms 3-5 would implicitly assume a minimal degree of comparability among different descriptions of a same consequence. This would be in line with the growing body of authors who consider completeness over consequences not to be a sound requirement of rationality. With a partial order, our results on unframed descriptive structure would also show that, if the preferences of the d_{ma} are not complete, they are still complete within classes of consequentially equivalent descriptions.

weak commutativity) and between the order of the concatenations of described consequences ($(x_1 \circ x_3) \circ x_2 \sim x_1 \circ (x_3 \circ x_2)$ by weak associativity). ■

Axiom 5 states that the d_{ma} is indifferent between a described consequence and the concatenation of this described consequence with itself. Formally, since the concatenation of a described consequence with itself is a logical possibility, it is necessary to include it in the basic setting. Intuitively, whether repetitions of described consequences do or do not increase informational content is an open question. For the d_{mo}, repeating the same information does not increase the informational content about the consequence. But the very fact of an information being repeated might impact the informational content about the consequence from the d_{ma}'s point of view, rationally or not. In any case, notice that such repetitions are not involved in the framing effects we are interested in, and there has been no empirical test of this to our knowledge.[9]

5.1.3 Framed descriptive structure with description invariance

The following axiom (also implicit in the standard model), together with the basic ones presented above, make our framework equivalent to the standard model, i.e., preference relations can only hold between consequences but not between different descriptions of the same consequences, among which the d_{ma} is necessarily indifferent.

Axiom 6 (Independence of common description). *If $x \circ z$ and $y \circ z$ are defined, then $x \sim y$ iff $x \circ z \sim y \circ z$.*

[9] A generalization of our framework should indeed dispense with the idempotency axiom as it is clear from both cognitive and social psychology that d_{ma}s are anything but indifferent to informational repetitions and redundancies (Cacioppo and Petty 1989; Judd and Braeur 1995; Moons et al. 2009; Kuhl and Anderson 2011; Mulligan and Peterson 2013; English and Visser 2014). Technically, such weakening might not be too difficult: the underlying algebraic structure associated to a consequence will no longer be a band but a commutative semigroup. Empirically, we did conduct some informal classroom experiments on the Asian Disease to check whether some of our axioms were or not respected. For the idempotency axiom, the results are very sensitive to the empirical counterpart of the concatenation. Roughly, using just a coma without logical connectors to concatenate the repetition could change the preferences between this concatenation and just one instance of the generator (and with Tversky and Kahneman's original generators, the overall preference reversal goes in the same direction than in their); but the use of a logical connector that made it clearer that it was a repetition (e.g., "so that") did not change the preferences (with Tversky and Kahneman's original generators, the overall preference reversal disappeared). We are currently preparing lab experiments on framing effects along these lines.

The independence of common description (ICD) axiom states that d_{ma}'s indifference between two described consequences is not affected by concatenation with a common described consequence. And conversely, d_{ma}'s indifference between two concatenations sharing a common described consequence is not affected by cancellation of the common described consequence.

The following implication of ICD is the crucial one that prevents a formal account of the framing effects we are interested in. We already said that the standard model cannot (without *ad hoc* assumptions) capture that different descriptions of the same consequence can convey different information; it only captures that different descriptions of the same consequence can convey the same information. Thus, it cannot discriminate between the following three cases. First, when different information are conveyed by different descriptions, if the information of one description is included in the information of another description, then we can say that the latter description *absorbs* the former: the informational content of their concatenation is identical with the informational content of the latter (absorbing) description. Second, when the same information is conveyed by different descriptions, then these descriptions are *mutually absorbent*. And third, when the information conveyed by a description absorbs all possible other information conveyed by other descriptions, we say that the former description is *maximally absorbent*. Formally, $x \in X$ absorbs $y \in X$ if and only if, $x \circ y \sim x$; $x, y \in X$ are mutually absorbent if and only if, $x \circ y \sim x \sim y$; $x \in X$ is maximally absorbent if and only if, for all $y \in S$, $x \circ y \sim x$. ICD states that either there exists no absorbing description or that all descriptions are maximally absorbent. But together with Axioms 1-5, ICD implies that every description of the same consequence is maximally absorbent, so that all description of the same consequence are mutually absorbent. In other words, the standard model cannot discriminate relative absorbency among different descriptions of the same consequence.

Example 2. From Table 5.1, $x_1 \circ x_2$ and $x_3 \circ x_2$ can be defined by the d_{mo} to correspond to:

$x_1 \circ x_2 \equiv$ "200 people will be saved and 400 people will die",

$x_3 \circ x_2 \equiv$ "400 people will not be saved and 400 people will die".

ICD implies that if the d_{ma} is indifferent between these two concatenations of described consequences, then he is indifferent between the two described consequences "200 people will be saved" and "400 people will not be saved", *and conversely.* ■

When Axioms 1-6 hold, we are in a *framed descriptive structure* with *description invariance*, which is equivalent to the standard model's *unframed descriptive structure.*

Proposition 1. *Let X be a nonempty set, $\approx$ and $\succsim$ nontrivial binary relations on X, and $\circ$ a partial binary operation on X. Suppose that the quadruple $\langle X, \approx, \succsim, \circ \rangle$ satisfies axioms 1-6. Then*

$$\forall x, y \in X, \quad x \approx y \Rightarrow x \sim y.$$

All proofs are in the appendix.

5.1.4 Framed descriptive structure with description-dependence

We will propose two ways of characterizing a *framed descriptive structure* with *description-dependence* in order to suitably represent the framing effects we are interested in, i.e., two ways our model can depart from the standard model's unframed descriptive structure. This subsection presents what is common to both ways; the next one presents the two characterizations and explains their differences. We keep the basic axioms and propose to weaken the ICD axiom in the same way for both characterizations. The ICD axiom can be decomposed in two parts:

Axiom 6.1 (Weak substituability) *If $x \circ z$ and $y \circ z$ are defined, then $x \sim y$ implies $x \circ z \sim y \circ z$.*

Axiom 6.2 (Weak simplifiability) *If $x \circ z$ and $y \circ z$ are defined,, then $x \circ z \sim y \circ z$ implies $x \sim y$.*

Example 2 above already illustrated the empirical interpretation of weak simplifiability, weak substituability was only implicitly illustrated through the "and conversely" clause. Weak substituability is the "if" direction of ICD: if the d_{ma} is indifferent between two informational

contents, then the simultaneous adjunction of another (consequentially equivalent) informational content preserves indifference. Weak simplifiability is the "only if" direction of ICD: if the d_{ma} is indifferent between two informational contents conveyed by two descriptions of the same consequence, then the retrieval of a common informational content preserves indifference between the remaining informational contents. It is this latter part that we drop to give a formal account of framing effects. To see why we drop weak simplifiability, consider the following example.

Example 3. From Table 5.1, $x_1 \circ (x_1 \circ x_2)$ and $x_2 \circ (x_1 \circ x_2)$ can be defined by the d_{mo} to correspond to:

$x_1 \circ (x_1 \circ x_2) \equiv$ "200 people will be saved and also 200 people will be saved and 400 people will die",

$x_2 \circ (x_1 \circ x_2) \equiv$ "400 people will die and also 200 people will be saved and 400 people will die".

Firstly, the basic axioms together with weak substituability imply that the d_{ma} is indifferent between these two described consequences.[10] Such indifference here is plausible, both intuitively because the concatenations can be seen as clarifying the informational content implied by either x_1 or x_2 and empirically because experimental results on the Asian Disease with the sure thing being completely described indeed does not reveal preference reversals (cf. previous chapter, §4.2.2).

Secondly, if weak simplifiability were to hold, it would further imply that the d_{ma} is indifferent between x_1 and x_2, i.e., between "200 people will be saved" and "400 people will die". Recall that, as explained in the previous chapter, this is descriptively inaccurate from the classical results of Tversky and Kahneman (1981); and normatively questionable from the results of Sher and McKenzie (2008) on informational leakage. We have therefore good reasons to drop weak simplifiability. ■

Without the weak simplifiability axiom, the result from the previous subsection does not hold anymore: the d_{ma} is no longer necessarily indifferent between different descriptions of the same

[10] This is so because by weak idempotence, $x_1 \circ x_1 \sim x_1$ and $x_2 \circ x_2 \sim x_2$. Weak substituability implies that $(x_1 \circ x_1) \circ x_2 \sim x_1 \circ x_2$ and $x_1 \circ (x_2 \circ x_2) \sim x_1 \circ x_2$. Rearranging by weak associativity, weak commutativity and transitivity of $\sim$ imply that $x_1 \circ (x_1 \circ x_2) \sim x_2 \circ (x_1 \circ x_2)$.

consequence. Consequently, he may exhibit preferences within a d_{mo}'s equivalence class, e.g., it is possible that, within $[x]$, $x_1 \succ x_2$ (or the reverse). It is because such cases of *description-dependence* are representable when axioms 1-5 and 6.1 hold (i.e., when 6.2 is dropped) that we say that axioms 1-5 and 6.1 characterize a framed descriptive structure with description-dependence. In this structure, it is possible that different descriptions of the same consequence can convey different information, i.e., all descriptions are not necessarily maximally absorbent anymore, especially generators, and the d_{ma} is not necessarily indifferent among them, he can prefer one description to another. But some indifference relations can still be implied.

Proposition 2. *Let X be a nonempty set, $\approx$ and $\succsim$ nontrivial binary relations on X, and $\circ$ a partial binary operation on X. Suppose that the quadruple $\langle X, \approx, \succsim, \circ \rangle$ satisfies axioms 1-5 and axiom 6.1. Then*

$$\forall x, y \in X, \quad x \text{ and } y \text{ are generated by the same generators } \Rightarrow x \sim y.$$

Moreover, for all $x \in X$, if $[x]_{\approx}$ is finitely generated by n generators, then $[x]_{\approx}$ has at most $2^n - 1$ equivalence classes under $\sim$.

Proposition 2 is weaker than proposition 1 in the sense that the d_{ma}'s indifference relation among described consequences is further constrained by their generators. Two different generators of the same consequence are not necessarily mutually absorbent and the d_{ma} is not necessarily indifferent between them. But two different descriptions generated by the same generators are necessarily mutually absorbent and the d_{ma} is necessarily indifferent between them. It also states that if the number of generators are limited within equivalence classes, then the structure of our objects of choice (described consequences) does not entail an infinity of indifference classes for a given consequence: the infinite number of described consequences does not allow an infinity of indifference classes (which would be absurd). The number of informational content possibilities is the number of distinct indifference classes, which is limited by the number of generators. This results can therefore be of practical use for experimenters who wish to ensure that they have tested (or at least considered) every logical possibility of different framings of the same decision problem.

Example 4. With x, y defined as in example 1, there exists, despite the infinite number of described consequences they can generate, at most 3 distinct indifference classes for the d_{ma}: $[x_1]_\sim$ generated by x_1, $[x_2]_\sim$ generated by x_2 and $[x_1 \circ x_2]_\sim$ generated by x_1 and x_2 (notice that they are all subclasses of one d_{mo} equivalence class, $[x]_\approx$ here). As already noted in the discussion of the idempotency axiom, it is an open question whether such repetitions of exactly the same descriptions of consequences (i.e., the empirical counterparts of these concatenations) does or does not increase informational content. This is not what is at stake in the framing effects we are interested in. What is at stake is the possibility of representing preferences between (but not within) elements of, respectively, $[x_1]_\sim$, $[x_2]_\sim$, and $[x_1 \circ x_2]_\sim$. ■

In the remaining of the chapter we work under two hypotheses. First, generators are finite within equivalence classes; we just argued for the plausibility of this hypothesis. Second, the number of consequences is finite (as in a given experimental setup, or in a given supermarket). This greatly facilitates the exposition of our next results.[11]

5.1.5 Tidy and untidy description-dependence

The following two main ways of characterizing description-dependence are worked out with axioms 1-5 and 6.1 holding. We wish to discriminate between cases of *tidy description-dependence* where the d_{ma}'s preferences over consequences are necessarily well-ordered, e.g., $[x]_\approx \succ [y]_\approx$ such that $x_1 \succ x_1 \circ x_2 \succ x_2 \succ y_1 \succ y_4 \succ y_2 \circ y_3$, from cases of *untidy description-dependence*, where preferences over consequences are not necessarily well-ordered, e.g., $[x]_\approx \succ [z]_\approx$ and $[z]_\approx \succ [x]_\approx$ (tantamount to a framing effect or preference reversal under certainty) as with $x_1 \succ x_1 \circ x_2 \succ z_3 \succ x_2$. The following two axioms characterize the discrimination between tidy and untidy description-dependence in so far as when they hold, it is necessarily the case that the d_{ma}'s

[11] The issue of the finite or infinite number of consequence is related to the issue of whether or not to include and Archimedean axiom in our framework. If such an axiom is included, we would have a less extensive structure in Krantz et al.'s sense, allowing us to explore utility representation which we do not explore in the subsection dedicated to that below, notably cases such as $u([a \circ b]_\sim) = u([a]_\sim) + u([b]_\sim)$ or $u([a \circ b]_\sim) = \lambda u([a]_\sim) + (1 - \lambda)u([b]_\sim)$. The empirical interpretation of such representations remains an open question. Formally, if we stay under the hypothesis that the number of consequences are finite, then we would need an infinite number of axiom (as in Fishburn 1970, p.42). That would however not be the case anymore if we worked under the hypothesis that the number of consequences are infinite. In any case, at a behavioral level, the empirical interpretations we have given of our framework under the finite consequence assumption and without an Archimedean axiom are enough to deal with all the framing effects presented in the previous chapter and to be presented in the last section of this one. Finally, with an infinite number of consequences and without an Archimedean axiom, we would have to deal with lexicographic preferences.

preferences are only tidily dependent to the descriptions of consequences, while when they are both dropped, his preferences can exhibit untidy description-dependence.

Axiom 7 (Non-triviality). *If $x \not\approx y$, then there exist $x' \in [x]_{\approx}, y' \in [y]_{\approx}$ such that $x' \not\sim y'$.*

Axiom 8 (Independence of common description across consequences). *If $x \not\approx z$, $x \circ y$ and $z \circ t$ are defined, then $x \succsim z$ iff $x \circ y \succsim z \circ t$.*

Non-triviality requires that for each pair of different consequences there is at least one description of each that actively affect the d_{ma}'s strict preference relation $\succ$. In other words, it ensures that there is at least one way of describing two different consequences so that the d_{ma} is not indifferent between them. Independence of common description across consequences (ICDAC) requires that for each pair of different consequences, neither concatenations nor unyoking of both change the d_{ma}'s preference ordering between them. Note that if ICD holds, ICDAC holds trivially, but not the converse. ICDAC together with non-triviality preclude cases of indifference between pairs of consequences implying indifference among all descriptions of both consequences, which is exactly what we want to avoid regarding strict framings of consequences.

Proposition 3. *Let X be a nonempty set, $\approx$ and $\succsim$ nontrivial binary relations on X, and $\circ$ a partial binary operation on X. Suppose that the quadruple $\langle X, \approx, \succsim, \circ \rangle$ satisfies axioms 1-5, 6.1, 7, and axiom 8. Then*

$$\forall x, y, z \in X, \text{ if } x \approx z \text{ and } x \succsim z, \text{ then } (x \succsim y \succsim z \Rightarrow y \approx x).$$

As said above, when both are dropped, we are within a framed descriptive structure with untidy description-dependence. Evidently, we could see further shades of tidiness and untidiness by weakening ICDAC. But we do not need to go so far to capture the strict framings of consequences we are interested in.

5.2 Utility representation and choice function

5.2.1 Representation theorems

The previous results can be translated in terms of utility representations. When only axioms 1-2 or the basic axioms hold, under the finiteness assumptions, there exists a real-valued function u on X, which represents d_{ma}'s preferences, that is, for all $x, y \in X$

$$x \succsim y \Leftrightarrow u(x) \geq u(y).$$

To see this, note that the finiteness assumptions imply that X is countable. Hence, representation in this case is a basic result. With the basic axioms, the image of X by u, denoted by $u(X)$, is at most countable. The results of propositions 1 and 2 imply that $u(X)$ is finite when we add ICD or just weak substituability. In the latter case, we are within a descriptive structure with untidy description-dependence, and the utilities of a given consequence lie in a bounded set, hence are contained in a finite interval that may overlap with other consequences' finite intervals of utilities. Within a descriptive structure with tidy description-dependence, i.e., with axioms 7-8 added, there is no such overlap. The following proposition summarize the relevant properties of the representation theorems (which are practically trivial to prove due to our finiteness assumptions, see the appendix):

Proposition 4. *Let $\langle X, \approx, \circ \rangle$ be a descriptive structure such that $X/\approx$ is finite and $[x]_\approx$ is finitely generated for all $x \in X$. For a binary relation $\succsim$ on X the following four statements hold:*

(i) *If $\succsim$ satisfies axioms 1 and 2 (and possibly one or more axioms among 3-5), then there exists an utility function on X that has an at most countable image;*

(ii) *If $\succsim$ satisfies axioms 1-6, then there exists an utility function on X that has a finite image. Moreover, for all $x, y \in X$, $a \approx b$ implies $u(x) = u(y)$;*

(iii) *If $\succsim$ satisfies axioms 1-5 and 6.1, then there exists an utility function on X that has a finite image. Moreover, for all $x, y \in X$, if x and y are generated by the same generators, then $u(x) = u(y)$;*

(iv) *If in addition, to the above conditions, $\succsim$ satisfies axioms 7 and 8, then for all $x, y \in X$, such that $x \approx y$ and $x \succsim y$, for all $z \in X$, $u(z) \in [u(x), u(y)]$ implies that $z \approx x$.*

5.2.2 Choice function

The goal now is to characterize the implications in terms of revealed preferences of the framework presented in the previous subsections. Yuval Salant and Rubinstein (2008) have provided a framework for what they call "choice with frames". However, none of their examples, applications and discussions consider the classical problem of description-dependence underlying Kahneman and Tversky's original framing effects. It may be thought that extending their framework through ours would allow to deal with violations of description invariance from a revealed preference perspective. This is not the case, and showing how allows to highlight the contribution of our framework.

Salant and Rubinstein extend the traditional approach in terms of choice function on the set of available alternatives A (i.e., a choice problem) from the set of alternatives X by constructing choice functions on pairs (A, f), where f is one frame from a set of frames F. Their framework can be extended to deal with framing effects due to different descriptions of the choice problem as follows. Extend the relation $\approx$ on the set of alternatives to the set of choice problems, denote this extended relation by $\approx_{\mathcal{P}}$. Define the pair $([A]_{\approx_{\mathcal{P}}}, f)$ where $[A]_{\approx_{\mathcal{P}}}$ is a standard choice problem with consequences and f indicates the description under which this choice problem is presented to the d_{ma}. Consider a descriptive structure without concatenation $\langle X, \approx \rangle$ such that X is finite (or equivalently, $X/\approx$ and $[x]_{\approx}$ are finite). Let $\mathcal{P}^*(X)$ be the set of all nonempty subsets of X and $\mathcal{P}^*(X/\approx)$ be the set of all nonempty subsets of $X/\approx$. For nonempty subsets of X, we will use notation $A, B, \ldots$, and for nonempty subsets of $X/\approx$, we will use notation $I, J, \ldots$. The relation $\approx_{\mathcal{P}}$ on $\mathcal{P}^*(X)$ is defined as: for all $A, B \in \mathcal{P}^*(X)$,

$$A \approx_{\mathcal{P}} B \text{ iff } (\forall x \in A, \exists y \in B, x \approx y) \text{ and } (\forall y \in B, \exists x \in A, y \approx x).$$

It is readily seen that the relation $\approx_{\mathcal{P}}$ on $\mathcal{P}^*(X)$ is an equivalence relation. More precisely, $\approx_{\mathcal{P}}$ extends the d_{mo}'s perspective on consequential equivalence in choice problems. A *described*

choice problem is a set of descriptions of alternatives. We can illustrate the previous definitions with the following example.

Example 5. Consider three consequences. One inspired from Levin and Gaeth (1988): eating ground beef $[G]$ described either as "25% fat ground beef" g_1 or "75% lean ground beef" g_2. The two other consequences are eating chicken C and eating beans B, for each of which we consider only one description, respectively c and b, which need not to be defined for the purpose of the example. Choosing between ground beef and beans is a choice problem $\{[G], [B]\}$ that can give rise to two different described choice problems: $\{g_1, b\}$ and $\{g_2, b\}$. From the d_{mo}'s perspective, they are two descriptions of the same choice problem: $\{g_1, b\} \approx_{\mathcal{P}} \{g_2, b\}$ because $g_1 \approx g_2$ and trivially $b \approx b$. ■

Moreover, to a choice problem we can associate all its descriptions, in particular there is a natural correspondence between $\mathcal{P}^*(X)/\approx_{\mathcal{P}}$ and $\mathcal{P}^*(X/\approx)$.

Proposition 5. *Let X be a set and $\approx$ an equivalence relation on X. There exists a bijection r from $\mathcal{P}^*(X/\approx)$ to $\mathcal{P}^*(X)/\approx_{\mathcal{P}}$.*

The above proposition justifies the following definition: an extended choice problem is a pair $([A]_{\approx_{\mathcal{P}}}, f)$ where $[A]_{\approx_{\mathcal{P}}}$ is a standard choice problem with consequences and f is a described choice problem, that is, any subset $A \subset X$ that belongs into $[A]_{\approx_{\mathcal{P}}}$.[12]

Following Salant and Rubinstein's approach, an *extended* choice function c assigns a chosen consequence $[x]_{\approx}$ to every pair $([A]_{\approx_{\mathcal{P}}}, f)$. Given a weak order $\succsim$ on X, the induced extended choice function $c_{\succsim}$ is defined by

$$c_{\succsim}([A]_{\approx_{\mathcal{P}}}, f) = \{[x]_{\approx} \in X/\approx \mid \exists x' \in [x]_{\approx} \cap f,\ x' \succsim y,\ \forall y \in f\}.$$

If X is finite, this set is necessarily nonempty. To see how their approach fails to capture framing effects violating description invariance, we propose the following counter example. From

[12] More formally, a standard choice problem with consequences is a subset $I \subseteq X/\approx$. By proposition 5, there exists an unique $[A]_{\approx_{\mathcal{P}}}$ such that $I = r^{-1}([A]_{\approx_{\mathcal{P}}})$, consequently with a slight abuse of notation and for the sake of clarity we denote a standard choice problem with consequences by $[A]_{\approx_{\mathcal{P}}}$ instead of I.

the previous example, suppose that the d_{ma} likes meat very much and derives more utility from what looks healthy, so that he prefers the ground beef described as "75% lean" over chicken over beans over the ground beef described as "25% fat", i.e., he exhibits the following profile of preferences: $g_2 \succ c \succ b \succ g_1$. Hence his preferences are untidy, there is no possible weak order on $X/\approx$. Furthermore the order on the described consequences violates Salant and Rubinstein's property ensuring the existence of an equivalence between an extend choice function and a weak order on X. They call this property γ^+-extended, which, adapted to deal with different descriptions of choice problems, can be stated as follow: if $c([A]_{\approx_\mathcal{P}}, f) = [x]_\approx$, $c([B]_{\approx_\mathcal{P}}, g) = [y]_\approx$ and $[y]_\approx \in [A]_{\approx_\mathcal{P}}$, then there exists a described choice problem h such that $c([A]_{\approx_\mathcal{P}} \cup [B]_{\approx_\mathcal{P}}, h) = [x]_\approx$. To illustrate a violation, note that according to d_{ma} preferences: $c(([B],[G]),(b,g_1)) = [B]$, $c(([C],[G]),(c,g_2)) = [G]$ and $[G] \in ([B],[G])$. However, $c(([B],[C],[G]),(b,c,g_1)) = [C]$ and $c(([B],[C],[G]),(b,c,g_2)) = c(([B],[C],[G]),(b,c,g_1,g_2)) = [G]$, hence $[B] \notin c(([B],[C],[G]),h)$ for all description h of the choice problem $([B],[C],[G])$.

To bypass the weaknesses of γ^+-extended to deal with descriptions of choice problems we propose the following modification:

γ^+-extended for descriptions: If $[x]_\approx \in c([A]_{\approx_\mathcal{P}}, f)$, $[y]_\approx \in c([B]_{\approx_\mathcal{P}}, g)$, $[x]_\approx \cap f$ is a singleton, and $[y]_\approx \cap g \subseteq f$, then for all described choice problem $h \in [A]_{\approx_\mathcal{P}} \cup [B]_{\approx_\mathcal{P}}$ such that $[x]_\approx \cap f \subseteq h \subseteq f \cup g$, it holds that $[x]_\approx \in c([A]_{\approx_\mathcal{P}} \cup [B]_{\approx_\mathcal{P}}, h)$.

This modification of γ^+-extended states roughly that if we are under the 'same frame' (e.g., *either* under the save frame *or* under the die frame in the Asian Disease), then there are no preference reversals. With this modification, Salant and Rubinstein's framework can deal with all preference reversals violating description invariance. However, the weak order generated by an extended choice function is not necessarily unique, preventing an account of tidy description-dependence.

Proposition 6. *Let X be a finite choice set of described consequences such that $X/\approx$ is not trivial and c an extended choice function. Then, c is an extended choice function that satisfies γ^+-extended for descriptions if, and only if, there exists a weak order $\succsim$ on X (not necessarily*

unique) such that $c = c_{\succsim}$.

The non uniqueness of the generated weak order makes it impossible to represent the modal preference of Gaeth and Levine's d_{ma}s, i.e., $g_2 \succ g_1$ if there are no other consequences between them. For instance, the following two profiles of preferences generate and are generated by the same extended choice function whereas the behavior is indeed not the same: $g_2 \succ g_1 \succ b$ and $g_1 \succ g_2 \succ b$. That is, their framework can represent $g_2 \succ g_1$ only by transitivity if $g_2 \succ c \succ g_1$.

To sum up, Salant and Rubinstein's framework, when extended to deal with descriptions of choice problems, does not allow to discriminate between the unframed descriptive structure of the standard model and the framed descriptive structure with tidy description-dependence of our framework. That is, if γ^+-extended for descriptions is satisfied we are necessarily in a case of untidy description-dependence if framing effects are observed, but if no such framing effects are observed, then we cannot know whether we are in a case of description invariance or tidy description-dependence.

In order to discriminate between tidy and untidy description-dependence in terms of revealed preferences, we propose to translate all our axioms in terms of choice functions. Let X be a set of described consequences, a choice function for X is a correspondence c from the set of all finite and nonempty subsets of X to X such that $\emptyset \neq c(S) \subseteq S$ for all $S \subseteq X$. The following conditions are translations of our axioms on preferences to be imposed on c:

Condition 1 (Descriptive structure) $\langle X, \circ, \approx \rangle$ *satisfies Definition 1.*[13]

Condition 2 (Contraction consistency and Expansion consistency (Sen 1971))

- *Contraction consistency: If* $x \in c(S)$ *and* $x \in T \subseteq S$*, then* $x \in c(T)$*;*
- *Expansion consistency: if* $x, y \in c(S)$*,* $S \subseteq T$*, and* $y \in c(T)$*, then* $x \in c(T)$*.*

Condition 3 (Commutativity consistency) *If* $x \circ y \in c(S)$*, then* $x \circ y, y \circ x \in c(S \cup \{y \circ x\})$.

Condition 4 (Associativity consistency) *If* $(x \circ y) \circ z \in c(S)$*, then* $(x \circ y) \circ z, x \circ (y \circ z) \in c(S \cup \{x \circ (y \circ z)\})$.

[13] Axioms 1 does not change because it does not involve the d_{ma}'s choice function.

Condition 5 (Idempotency consistency) *If $x \in c(S)$, then $x, x \circ x \in c(S \cup \{x \circ x\})$.*

Condition 6.1 (Substituability consistency) *If $x, y \in c(S)$ and $x \circ z, y \circ z \in T$, then $x \circ z \in c(T)$ iff $y \circ z \in c(T)$.*

Condition 6.2 (Simplifiability consistency) *If $x \circ z, y \circ z \in c(S)$ and $x, y \in T$, then $x \in c(T)$ iff $y \in c(T)$.*

Condition 7 (Non-triviality) *If $x \not\approx b$, then there exist $S \subset X$, finite, $x' \in S \cap [x]_{\approx}$, $y' \in S \cap [y]_{\approx}$ such that $x' \in c(S)$ and $y' \notin c(S)$.*

Condition 8 (ICDAC consistency) *Suppose $x \not\approx z$, $x \approx y$ and $z \approx t$.*

- *If $x, z \in S$, $x \in c(S)$, then there exists $T \subset X$, finite, such that $x \circ y, z \circ t \in T$ and $x \circ y \in c(T)$;*
- *if $x \circ y, z \circ t \in S$, $x \circ y \in c(S)$, then there exists $T \subset X$, finite, such that $x, z \in T$ and $x \in c(T)$.*

These conditions characterize the behavioral dimension of description invariance and its violations discussed previously. The revealed preferences will satisfy *description invariance* if conditions 1-5 and conditions 6.1 and 6.2 are satisfied, *tidy description-dependence* if conditions 1-5 and conditions 6.1, 7 and 8 are satisfied (but not condition 6.2), and *untidy description-dependence* if conditions 1-5 and condition 6.1 are satisfied (but not conditions 6.2 and 8).

Proposition 7. Let X be a set of described consequences and c be a choice function for X. Then, c is a choice function that satisfies condition C_i if, and only if, there exists a binary relation $\succsim$ on X such that $c = c_{\succsim}$ and $\succsim$ satisfies axiom A_i.

5.3 Discussion: application, extensions and related literature

To show how our framework can be used to deal with framing effects and decision problems more generally, we apply it to four sets of phenomena, namely to some cases of (5.3.1) framing of

consequences under risk, (5.3.2) framing of acts and of contingencies (that we discuss altogether and under both risk and uncertainty), (5.3.3) framing under certainty and (5.3.4) individuation of consequences in the Allais paradox. For each of these, we contrast our framework with existing formal contributions in economics and show when the compatibility between both allows for non-trivial extensions.

5.3.1 The framing of consequences under risk

Given the illustrative example used throughout the construction of our framework, its most natural application is to the original Asian Disease problem *under risk* (remember that the framework was built under certainty hence could not deal with the classical Asian Disease problems). The extension consists in using our framework within existing models of decision under risk that respect an axiom of weak coalescing, which is tantamount to working with these models on our choice set[14]. For instance, models in expected utility theory, rank-dependent utility theory and cumulative prospect theory respect that condition. Let X be a nonempty set of described consequences (space of risk-less outcomes) and $\Delta(X)$ be the set of all simple distributions on X. The descriptive structure is now $\langle \Delta(X), \approx, \circ \rangle$. This descriptive structure induces another descriptive structure $\langle X, \approx_X, \circ_X \rangle$ on described consequences[15]. It follows that we can use the utility of described consequences in theories that respect weak coalescing. Our results ensure that this replacement does not alter the axioms underlying these theories. It is worth noting that under expected utility, the implied risk attitudes are dependent on the descriptions of the consequences, and not the consequences itself, which is the whole point of Asian Disease problems in the first place. Formally, using the notation from Table 5.1, the modal preferences in the original Asian Disease can be represented straightforwardly in expected utility theory (defined over our choice set) as:

[14] Weak coalescing holds if for all $x, y \in X$ such that $x \sim y$, $(x, p; y, q; z, 1-p-q) \sim (x, p+q; z, 1-p-q)$. This property is stronger than coalescing with $x = y$.

[15] Another way of extending our framework under risk would be to follow Giraud (2004a, pp.7-11): Start from $\langle X, \approx, \circ \rangle$ and extend it to $\Delta(X)$. It is natural to extend $\approx$ on X as follows: for all $l, m \in \Delta(X)$, $l \approx m$ iff $(\forall x \in s(l), \exists y \in s(m), x \approx y, l(x) = m(y))$ and $(\forall y \in s(m), \exists x \in s(l), y \approx x, m(y) = l(x))$, and $\circ$ pointwise. This approach indeed works for the framing effects we are interested in, but leave aside logically equivalent lotteries (for instance the ones in splitting effects, i.e., violating coalescing, see Birnbaum 2005). To avoid this drawback, $\approx$ mus be extended on $\Delta(X)$ by: $l \approx m$ iff the induced lotteries on consequences are equal. In this case, the definition of $\circ$ becomes more tedious.

$$u(x_1) > \frac{1}{3}u(y_1) + \frac{2}{3}u(z_1)$$
$$\frac{1}{3}u(y_2) + \frac{2}{3}u(z_2) > u(x_2)$$

These inequalities can be rationalized by one weak order in the sense that $u(.)$ is its numerical representation and that it generates and can be generated by a choice function as defined in the previous section. Furthermore, the same weak order can also rationalize all the other empirical regularities on the Asian Disease problem.

The same reasoning carries identically to other framing of consequences besides the Asian Disease. Two further illustrations are worth providing. One is DeMartino et al. (2006)'s where the two different frames have identical descriptions of the probabilistic options, they differ only by different descriptions of the sure options (cf. previous chapter). Formally, from the d_{mo}'s perspective, let $a_1 \equiv$ "Keep £20", $a_2 \equiv$ "lose £30" $\in [a]_{\approx}$, $b_1 \equiv$ "keep all" $\in [b]_{\approx}$ and $c_1 \equiv$ "lose all" $\in [c]_{\approx}$. The modal perspective from d_{ma}s can be represented straightforwardly in expected utility theory:

$$u(a_1) > \frac{1}{3}u(b_1) + \frac{2}{3}u(c_1)$$
$$\frac{1}{3}u(b_1) + \frac{2}{3}u(c_1) > u(a_2)$$

Notice the formal representation of the distinction between, on the one hand, a difference in descriptions of the same consequence (for the sure one), and, on the other hand, an identity of descriptions of the same consequence (for the probabilistic ones). The last illustration comes from McNeil et al. (1982) and has a different empirical structure than Asian Disease-like problems. d_{ma}s are presented with hypothetical choices between two ways of treating lung cancer, i.e., surgery vs. radiation therapy, in one of two frames. In one frame the cumulative probabilities are presented as descriptions of survival rates:[16]

[16]The exact empirical wording of the problem appears only in Tversky and Kahenman (1986).

> **Surgery**: Of 100 people having surgery 90 live through the post-operative period, 68 are alive at the end of the first year and 34 are alive at the end of five years.
> Radiation Therapy: Of 100 people having radiation therapy all live through the treatment, 77 are alive at the end of one year and 22 are alive at the end of five years.

In another frame, they are presented as descriptions of death rates:

> Surgery: Of 100 people having surgery 10 die during surgery or the post-operative period, 32 die by the end of the first year and 66 die by the end of five years.
> **Radiation Therapy**: Of 100 people having radiation therapy, none die during treatment, 23 die by the end of one year and 78 die by the end of five years.

In this framing problem, the objects of choice are streams of dated binary lotteries. In each lottery, the binary consequences are mutually exclusive (i.e., being dead or alive) and from the d_{mo}'s perspective, presenting only one of the two branches (consequence plus associated probability) allows to infer the other branch and thus the complete lottery. Furthermore, we can define a logical equivalence between objects of choice over time (as was for the extension of our framework for decision under risk). Under the standard assumption of inter-temporal separability, we can focus only on lotteries independently of their place in the stream.

Formally, let

$$\left.\begin{array}{l} a_1 \equiv \text{“}\frac{90}{100}\text{ chances of living through the post-operative period”} \\ a_2 \equiv \text{“}\frac{10}{100}\text{ chances of dying during the post-operative period”} \end{array}\right\} \in [a]_{\approx}$$

$$\left.\begin{array}{l} b_1 \equiv \text{“}\frac{68}{100}\text{chances of living in one year”} \\ b_2 \equiv \text{“}\frac{32}{100}\text{chances of dying in one year”} \end{array}\right\} \in [b]_{\approx}$$

$$\left.\begin{array}{l} c_1 \equiv \text{“}\frac{34}{100}\text{chances of living in five years”} \\ c_2 \equiv \text{“}\frac{66}{100}\text{chances of dying in five years”} \end{array}\right\} \in [c]_{\approx}$$

$$\left.\begin{array}{l} d_1 \equiv \text{“}\frac{100}{100}\text{chances of living through the post-operative period”} \\ d_2 \equiv \text{“}\frac{0}{100}\text{chances of dying through the post-operative period”} \end{array}\right\} \in [d]_{\approx}$$

$$\left.\begin{array}{ll} e_1 \equiv \text{“}\frac{77}{100}\text{chances of living in one year”} & , \\ e_2 \equiv \text{“}\frac{23}{100}\text{chances of dying in one year”} & \end{array}\right\} \in [e]_{\approx}$$

$$\left.\begin{array}{ll} f_1 \equiv \text{“}\frac{22}{100}\text{chances of living in five years”} & , \\ f_2 \equiv \text{“}\frac{78}{100}\text{chances of dying in five years”} & \end{array}\right\} \in [f]_{\approx}$$

The modal perspective from d_{ma}s can be represented straightforwardly with the subscripts 0, $1y$ and $5y$ denoting respectively the postoperative period, one year after the operation and five years after the operation:

$$u_0(a_1) + u_{1y}(b_1) + u_{5y}(c_1) > u_0(d_1) + u_{1y}(e_1) + u_{5y}(f_1)$$
$$u_0(d_2) + u_{1y}(e_2) + u_{5y}(f_2) > u_0(a_2) + u_{1y}(b_2) + u_{5y}(c_2)$$

Again, the results of our framework are enough to ensure that there is a binary relation of preference underlying u and that it generates and can be generated by a choice function.

Let us now contrast our framework with other formal accounts of framing effects under risk. The main difference with all of these accounts is that we considered a richer set of behavioral data from the psychology literature to guide our axiomatic constructions. Among these accounts, the closest from our framework is Giraud's (2004a; b; 2005). We were indeed inspired from what he calls a "normative" equivalence relation on the choice set, which is close to the perspective of the d_{mo} in our framework. However, we did not necessarily identify the d_{mo}'s perspective with a normative one that the d_{ma} would like to take and we characterized their relations through an algebraic structure on descriptions. Hence we were able to decompose our invariance axiom (ICD) and characterize how a utility and choice function still exist when this axiom is weakened. That is, we went further than imposing the axiom and forcing a choice between taking it to be true or false[17]. This considerably simplifies applications, especially because Giraud cannot formalize a direct dependence of preferences on descriptions without interpreting it as a failure of the d_{ma} to recognize that different descriptions are about the same consequences.

[17] We thank Giraud for explaining to us these differences in these terms

To bypass this shortcoming, he gives an account of the original Asian Disease which is more complex than ours, where the preference reversal is explained through a nonadditive decision weight function, making the d_{ma}'s beliefs directly dependent on descriptions and preferences indirectly so when revealed (see Giraud 2004a; b). Finally, he makes more assumptions about the underlying psychological processes of the d_{ma} on which we are agnostic, and while we narrowly focused on the role of descriptions, his framework is also built to deal with violations of context independence (i.e., preference reversals triggered by expansion and/or contraction of the choice set) and of procedure invariance (i.e., preference reversals triggered by different elicitation methods). Hence, because we believe that our frameworks are compatible, incorporating our choice set in his approach could make it more tractable while expanding the explanatory and empirical scope of ours.

The few other formal accounts of framing effects in the literature have all been conducted in frameworks that are less close to the standard one than Giraud's and ours. Nevertheless under one form or another, the differences just discussed resurface.

Take for instance Blume et al. (2013). They account for framing effects within their "constructive decision theory" framework where objects of choice are syntactic descriptions of events in a simple programming language. They show how this language naturally formalizes the two frames of McNeil et al. (Surgery vs. Radiation) as two logically non-equivalent syntactic descriptions. Description invariance is not derived and then weakened but imposed by a subset of the syntactic descriptions that describes the same events according to the d_{ma}'s *beliefs*. This allows to represent both d_{ma} revealing framing effects and those who don't (simply by putting the two syntactic descriptions of the two frames in different subsets or by regrouping them within the same subset, respectively).

Or take Gold and List (2004). They use the resources of predicate logic to construct a choice set in which the elements are either "target" propositions about the d_{ma}'s preferences or various "background" propositions about the decision situation. Framing effects are formalized as arising from a sequential process whereby *beliefs* about two different sets of background propositions can influence contradictory beliefs about the same target proposition, e.g., the dm_a believes that it is true that he prefers the sure program over the probabilistic one and the d_{ma} believes that it is false that he prefers the sure program over the probabilistic one (see pp.267-268). They

make description invariance formally explicit but not as quite straightforwardly as in our and others' frameworks.

Take finally Bourgeois-Gironde and Giraud (2009). They analyze framing effects through the framework of Bolker-Jeffrey expected utility where preferences are over propositions. They make explicit the implicit axiom of invariance in this framework by showing how it is hidden both in the mathematical structure of the Boolean algebra with which the choice set is constructed and in its interpretation as a set of propositions. Once made explicit, they however do not weaken it, but propose to "bypass" it (p.390) by introducing a new set of information ("good news" and "bad news") to be associated with each events, hence refining the d_{ma}'s *beliefs* about the latter. Following McKenzie and co-authors (see the previous chapter, §4.3.1), framing effects are thus rationalized by refining invariance to the informational (and not only logical) structure of decision problems.

Among these three contributions, the closest to ours in spirit is Bourgeois-Gironde and Giraud (2009) because of their inspiration from McKenzie and co-authors' results to motivate an account of framing effects as not necessarily irrational. They however do not consider the further results from psychology that motivated our formalization of a direct dependence of preferences on descriptions. Bacharach (2003) provides an analysis of framing effects under risk in this spirit, but it is quite informal and less motivated by behavioral data (concerning the direct dependence of preferences on descriptions).[18]

Summing up, the main difference between ours and others' frameworks is the axiomatic approach through which we make description invariance explicit and then weaken it to allow for a direct dependence of preferences on descriptions. The main advantages of this approach are its simplicity (both of the framework itself and in its application) and its fit with existing behavioral data on framing effects in psychology. Notice that these data include not only Kahneman and Tversky's original ones, but also the ones from other psychologists presented in the previous chapter. One of the points of their early experiments was to demonstrate framing effects by framing consequences, acts and events *independently* of each others. This threefold empirical

[18] Both Ryan (2005) and Lanzi (2011) claim to provide a framework to formalize framing effects. Their accounts do not seem very tractable; in any case they do not provide concrete applications to framing effects showing how the formalization should work (they do so for other phenomena). Lerner (2014) provides an account of framing effects through intensional logic. Though she illustrates how her framework formalizes the Asian Disease, it is again not very tractable.

and conceptual distinction seems lost in others' contributions, all of which introduce events and beliefs in the formalization of problems such as the Asian Disease and other framing effects under risk where there are no events and where only the consequences are framed[19]. What does our framework have to say on beliefs and the framing of events and acts?

5.3.2 The framing of acts and events, under risk and uncertainty

To deal with this question, we propose to contrast our framework with Ahn and Ergin (2010). The latter is primarily designed to account for framing effects due to the framing of events in cases such as the following (ibid, p.655). Consider these two health insurance contracts stipulating the consequences (i.e., the amounts of money in their left sides) occurring contingently upon the realization of events (i.e., described in their right sides):

$$\begin{pmatrix} \$500 & \text{surgery} \\ \$100 & \text{prenatal care} \\ \vdots & \vdots \end{pmatrix} \quad \begin{pmatrix} \$500 & \text{laminotomy} \\ \$500 & \text{other surgeries} \\ \$100 & \text{prenatal care} \\ \vdots & \vdots \end{pmatrix}$$

Notice that "laminotomy" and "other surgeries" are both included in "surgery", and the union of the two former is equal to the latter. Hence we have two different descriptions of the same contract. Ahn and Ergin's axiomatic approach consists in making explicit the axiom of invariance to such redescriptions that is implicit in the standard model, and then weaken it to provide what they call a "partition-dependent expected utility" theory.

Our framework cannot deal directly with such framing effects. It could if properly extended for acts as objects of choice instead of consequences. Such extension would be in line with Ahn and Ergin's framework because, due to our finiteness assumptions, we would need to weaken description invariance with a finite state-space à la Anscombe and Aumann (1963) as they do, i.e., not within an infinite one à la Savage)[20]. In a way, Ahn and Ergin's elegant results would make it easy for us to carry such an extension, which may not be trivial in the sense of accounting for at least two further types of framing phenomena for which they cannot in

[19] With the exception of Bacharach (2003, p.66), who insists that "*Framing is logically prior to believing.*"

[20] Or to do so in Savage's framework, we could work our extension trough Gul's (1992) modification of the former for finite state-space.

their framework, namely those due to different descriptions (1) of the same consequence or (2) of the same events when the redescription is "pure", i.e., not made through the packing or unpacking of an inclusion relation. The two following examples respectively illustrate each case and are inspired from the field studies and lab experiments of Sinaceur et al. (2005) about the impacts on various cognitive and behavioral measures of different descriptions of the same disease either with the ordinary label "Mad Cow" or with the scientific label "bovine spongiform encephalopathy" (BSE). Imagine that you ate beef and the consequences of that act are that you get the disease in the event that the U.K. trades beef with France (abbreviated by "trades with U.K.") and that you don't get the disease otherwise. Here are two different descriptions of that act:

$$\begin{pmatrix} \text{get Mad Cow} & \text{trades with U.K.} \\ \text{don't get Mad Cow} & \text{no trades with U.K.} \\ \vdots & \vdots \end{pmatrix} \begin{pmatrix} \text{get BSE} & \text{trades with U.K.} \\ \text{don't get BSE} & \text{no trades with U.K.} \\ \vdots & \vdots \end{pmatrix}$$

Now imagine that you ate beef at Mike's place, and the consequences of that act is that you get sick in the event that Mike bought infected beef and don't get sick otherwise. Here are two different descriptions of that act:

$$\begin{pmatrix} \text{get sick} & \text{Mike bought Mad Cow beef} \\ \text{don't get sick} & \text{Mike bought healthy beef} \\ \vdots & \vdots \end{pmatrix} \begin{pmatrix} \text{get sick} & \text{Mike bought BSE beef} \\ \text{don't get sick} & \text{Mike bought healthy beef} \\ \vdots & \vdots \end{pmatrix}$$

It can be argued that Ahn and Ergin's impossibility to deal with this latter example is problematic in the sense that their model is devised to account for the effects of redescriptions of the same event, while it is less problematic that they cannot deal with the former because doing so is not their contention. On the other hand, we could deal with the former case quite straightforwardly because it is the kind of framing effects we designed our framework for, while the latter case would require more work. As we see it, a state-dependent account may be the sensible way to go here. We believe so because state-dependent utility finds a natural interpretation that is well in line with the work of McKenzie on information leakage which

motivated our framework. The interpretation is that, it is not the descriptions of the event *per se* on which the d_{ma}'s preferences are dependent, but on *the fact that an entity chose a given description* over other possible ones. Hence preference reversals due to 'who said what' could be formally represented in a neat way.

Finally, let us consider Tversky and Kahneman's (1981, p.454) framing of act under risk discussed in the previous chapter (though "act" is not to be taken in a Savagean sense because there are no events in their example). It is worth reproducing the decision problem here for convenience:

[First frame]

> Imagine that you face the following pair of concurrent decisions. First examine both decisions, then indicate the options you prefer.
>
> Decision (i). Choose between:
>
> **A**. a sure gain of \$240
>
> *B*. 25% chance to gain \$1000, and 75% chance to gain nothing
>
> Decision (ii). Choose between:
>
> *C*. a sure loss of \$750
>
> **D**. 75% chance to lose \$1000, and 25% chance to lose nothing

[Second frame]

> Choose between:
>
> *A*&*D*. 25% chance to win \$240, and 75% chance to lose \$760.
>
> **B**&**C**. 25% chance to win \$250, and 75% chance to lose \$750.

In this framing problem, the object of choice are lotteries that are either composed in the first frame or reduced in the second one. From the d_{mo}'s perspective, both forms of lotteries are equivalent, i.e., $(A, D), A\&D \in [AD]_{\approx}$ and $(B, C), B\&C \in [BC]_{\approx}$. The modal preference from d_{ma}s, i.e., $(A, D) \succ (B, C)$ and $B\&C \succ A\&D$, exhibit untidy description-dependence and hence violate ICDAC. However, to fully represent these choices in terms of utility functions on described consequences with their associated probabilities, we would need to incorporate our

framework in models allowing for violations of first-order stochastic dominance (see the ones presented by Birnbaum 2005).

5.3.3 Framing under certainty

In this subsection we briefly apply our framework to two famous cases of framing under certainty to illustrate possible interpretations of the d_{ma}-d_{mo} relation for the positive/normative issue, i.e., regarding the rationality or irrationality of framing effects. The first case is from Quattrone and Tversky (1988). It can be thought of as an analogue of the Asian Disease under certainty where the choice is between two public policies with different trade-offs between (un)employment and inflation:

> If **program A** is adopted, 5% of the work force would be unemployed, while the rate of inflation would be 17%.
> If program B is adopted, 10% of the work force would be unemployed, while the rate of inflation would be 12%.
>
> If program C is adopted, 95% of the work force would be employed, while the rate of inflation would be 17%.
> If **program D** is adopted, 90% of the work force would be employed, while the rate of inflation would be 12%.

In our framework, there are two straightforward ways to represent these modal preferences. Firstly, the d_{mo} is an economist who classes objects of choice in terms of their economic implications:

$$\left.\begin{array}{l} a_1 \equiv \text{“5\%of the work force would be unemployed, while the rate of inflation would be 17\%”} \\ a_2 \equiv \text{“95\%of the work force would be employed, while the rate of inflation would be 17\%”} \end{array}\right\} \in [a]_{\approx}$$

$$\left.\begin{array}{l} b_1 \equiv \text{“10\%of the work force would be unemployed, while the rate of inflation would be 12\%”} \\ b_2 \equiv \text{“90\%of the work force would be employed, while the rate of inflation would be 12\%”} \end{array}\right\} \in [b]_{\approx}$$

The modal perspective from d_{ma}s can be represented straightforwardly as $u(a_1) > u(b_1)$ and $u(b_2) > u(a_2)$. If the d_{mo} wish to take a normative perspective on these preferences, he can

either argue that they are rational because representable by a utility and choice functions, or irrational because they exhibit untidy description-dependence and hence violate ICDAC.

Secondly, the d_{mo} is a politician who classes objects of choice in terms of whether they look like good news or like bad news, respectively:

$$\left.\begin{array}{l} a_1 \equiv \text{“5\%of the work force would be unemployed, while the rate of inflation would be 17\%”} \\ a_2 \equiv \text{“90\%of the work force would be employed, while the rate of inflation would be 12\%”} \end{array}\right\} \in [a]_{\approx}$$

$$\left.\begin{array}{l} b_1 \equiv \text{“10\%of the work force would be unemployed, while the rate of inflation would be 12\%”} \\ b_2 \equiv \text{“95\%of the work force would be employed, while the rate of inflation would be 17\%”} \end{array}\right\} \in [b]_{\approx}$$

In this case, the modal perspective from the d_{ma}s can be represented straightforwardly as $u(a_1) > u(b_1)$ and $u(a_2) > u(b_2)$. If the d_{mo} wishes to take a normative perspective on these preferences he can argue that they are rational not only because they are representable as utility maximization and choice behavior, but also because they exhibit tidy description-dependence and hence do not violate ICDAC. Alternatively, the d_{mo} could argue for their irrationality because they exhibit tidy description-dependence and hence violate ICD.

The second case of framing under certainty is Tversky and Kahneman's (1981) lost bill/lost ticket example discussed in the previous chapter. Recall that, on the one hand, the two frames are about two different situations (not two descriptions of the same situation, i.e., it was a loose framing effects), and, on the other hand, beyond the lab this problem cannot really be interpreted as an external d_{mo} posing a decision problem to the d_{ma} (in the real world you would be both the d_{mo} and the d_{ma} and construct your own counterfactuals to make up your mind). Nevertheless, it is possible to capture both these differences in our framework, and again in two different but straightforward manner. Firstly, the d_{mo} is either an observing economist or a self of a multiple selves d_{ma}, both of whom class objects of choice by their consequences in terms of opportunity costs:

$$\left.\begin{array}{l} a_1 \equiv \text{“paying for a \$10 ticket after having lost a \$10 bill”} \\ a_2 \equiv \text{“paying for a \$10 ticket after having lost a \$10 ticket”} \end{array}\right\} \in [a]_{\approx}$$

In this case, the modal preference of d_{ma}s can be represented straightforwardly as $u(a_1) > u(a_2)$. If the d_{mo} wishes to take a normative perspective on these preferences, he can either argue that they are rational because representable by a utility function and a choice function, or irrational because they exhibit tidy description-dependence and hence violate ICD.

Secondly, the d_{mo} is a self of a multiple selves d_{ma} who classes objects of choice by their consequences in terms of their experienced utility, considering that the experience of being upset from having lost a ticket is different from the experience of of being upset from having lost a bill, respectively:

$$a_1 \equiv \text{“paying for a \$10 ticket after having lost a \$10 ticket”} \in [a]_{\approx}$$

$$b_1 \equiv \text{“paying for a \$10 ticket after having lost a \$10 bill”} \in [b]_{\approx}$$

In this case, the modal preference of d_{ma}s can be represented straightforwardly as $u(a_1) > u(b_1)$. If the d_{mo} wishes to take a normative perspective on these preferences, he can either argue that they are rational because representable by a utility function and a choice function, or that they are irrational because they exhibit untidy description-dependence and hence violate ICDAC.

By presenting the possibility of multiple interpretations from our framework we do not aim to take a relativistic perspective on the normative issues underlying framing effects. Instead, we believe that these issues are important and that the flexibility of our framework allows to make precise discussions of the conditions under which a given interpretation of observed preferences can be seen as rational or irrational. In that sense, though formally quite different, we take a position similar to Gilboa et al. (2010), i.e., that a formal framework can help organize rational discussions and even debates over normative issues underlying individual rationality in economics.

5.3.4 Individuation and redescription of consequences

In this subsection, we suggest that something very close to our framework was presupposed in a set of arguments constitutive of the debates around Allais' paradoxes. Recall that the classical

Allais paradox (a version of the so-called common consequence effect) goes as follows (Allais 1953, p.527; consequences are monetary, originally in French francs):

> **A**: The certainty of winning 100 millions
>
> B: 10% chance of winning 500 millions, 89% chance of winning 100 millions, and 1% chance of winning nothing

Notice that both A and B share a common consequence, namely "winning 100 millions", with certainty in A and with 89% chance in B. Thus both A and B share at least 89% of winning 100 millions (and A has an extra 11% chance for it). So, if we retrieve this common consequence with the common probability of occurrence, the preference between A and B should not change to respect the independence axiom. But it does:

> C: 11% chance of winning 100 millions
>
> **D**: 10% chance of winning 500 millions and 90% chance of winning nothing

As reviewed by Broome (1991), most defenses of expected utility theory consisted in arguing that Allais' paradoxes are not paradoxes anymore when the consequences are redescribed to add information about, for instance, some feelings of disappointment if the decision maker does not win anything. Such redescriptions, e.g., "winning nothing" and "winning nothing and being disappointed" are then taken as different descriptions of *different* consequences, hence breaking the equivalence between the two initial pairs of problem, and making it impossible to observe an inconsistency from expected utility theory. Within our framework, this reasoning could be formalized as follows. The d_{mo} is a decision theorist who classes consequences in terms of their experienced utility, considering that the experience of being disappointed when winning nothing is different from the experience of winning nothing:

$$a_1 \equiv \text{“winning nothing”} \in [a]_{\approx}$$

$$b_1 \equiv \text{“winning nothing and being disappointed”} \in [b]_{\approx}$$

By contrast, those who argued against this interpretation held a position that can be formalized as :

$$\left.\begin{array}{ll} a_1 \equiv \text{“winning nothing”} & , \\ a_2 \equiv \text{“winning nothing and being disappointed”} & \end{array}\right\} \in [a]_{\approx}$$

It can be argued that the condition of possibility for this debate is the standard notion of a consequence, which, following Savage (1972 [1954], pp.13), is taken to mean "anything that may happen to the person", i.e., to the d_{ma}. This includes any "things, or experiences, regarded as consequences" (ibid, p.14). It may be asked, regarded as a consequence by whom? In standard decision theory, consequences are, in principle, regarded as such by d_{ma}s. But by virtue of revealed preference methodology, only the d_{ma}'s obervable choice behavior should be a primitive in the analysis which excludes his subjective perception of what counts as one consequence. Hence, in practice, consequences are regarded as such by a decision theorist guessing or postulating what a d_{ma} would count as a consequence. The first position in the debate on redescriptions in the Allais paradox shows that one drawback of this permissive notion of a consequence is that it prevents to account for the simple fact that one d_{ma} (or even one decision theorist for that matter) can entertain different perspectives on the *same* consequence. That is, it prevents to account for the recognition by the d_{ma} that different presentations of the same consequence (or "things") trigger different preferences (or "experiences"). Instead, it can be argued that we shall try to do justice to the subjectivity of a d_{ma} who recognizes that one and the same consequence can trigger different mental states in different situation, without conveniently postulating that the differences were "in fact" about different consequences. This is the second position in the debate, which was notably held by Sen (2002, chap.6 and chap.11 [both from 1986]) and Machina (1989, sect.6.6). Following these two authors, we argue that in order to do justice to the decision maker's values and hence make the related normative analysis easier, we shall not blend all the factually relevant observations into one single theoretical construct (here of "consequence"). The empirical stakes are just as high as the normative ones here, since too permissive a notion of a consequence makes any utility theory not amenable to empirical falsification *in principle*, a methodological position which is not quite the same as not being falsifiable *in practice*. Therefore, what underlies our whole framework is not the standard notion of a consequence, but rather what Machina (1989, p.1660) calls "the operational definition of a consequences". We do not see this as a drawback, as it can be argued that the whole of experimental economics (or at least its subpart that focuses on decision theory) does indeed not rely on the standard notion of a consequence. We do not see how framing effects could even be intelligible with that notion.

In short, we argue that different descriptions of one same consequence that can be described as such by a d_{mo} and recognized as such by a d_{ma} should be represented as such by the decision theorist (if he is not already the d_{mo}). By virtue of the formal discrimination of different descriptions of a same consequence, our framework allows to do this in a straightforward manner. With it, we have tried to shed some lights on a debate around the Allais paradox which is not a framing effect per se. Yet the underlying issues are not unconnected with framing effects (for a sustained argument in that direction, see Giraud 2004b, pp.146-7). Whether out framework can help to resolve deeper conceptual issues related to the permissive notion of consequence in decision theory (see Aumann 1987 [1971]; Savage 1987 [1971]) is an open question that we left for further work.

Conclusion: the decision maker and the decision modeler, together at last

In this chapter, we proposed an account, in the language of economics, of the framing effects violating description invariance that were presented in the previous chapter. This was made through an axiomatic framework in which the axiom of description invariance that is implicit in standard models could be made explicit and then weakened. We have seen that, to make it explicit, the traditional choice set first needs to be refined from objects of choice to their descriptions, and then description invariance can be characterized as an indifference relation between all the descriptions *and their possible concatenations* of a given object of choice. Furthermore, different weakening of description invariance can imply different different degrees of dependence to descriptions. We have characterized the implications in terms of choice, preference and utility of two such degrees, i.e., tidy description-dependence (preferences only within but not across consequences) and untidy description-dependence (preferences within and across consequences). Under tidy description-dependence, the Asian Disease and other well known framing effects can be formally accounted for rather straightforwardly by integrating our choice set in standard (as well as non-standard) models. For other framing effects, this is not so straightforward and more work is needed. This is notably the case for framing effects due to different descriptions of the same event, for which we should work on the d_{ma}'s beliefs. Because we have stated our

framework in terms of informational content (instead of merely 'descriptions'), such an extension is conceptually unproblematic. Notice that it is only through an extension to beliefs that a full characterization of Sher and McKenzie's informational leaking account discussed in the previous chapter will be possible. We have suggested a way to make this extension in line with the theory of speech act underlying the methodological perspective of this dissertation, namely through state-dependent preferences with states identified as a speech act from a d_{mo} .

To conclude, we would like to emphasize the possible mutual-influences between the methodological and philosophical approach taken in all the previous chapters and the formal one taken here. On the one hand, it is straightforward that the methodological and philosophical considerations of the previous chapter were the conditions of possibility of this one. Without the previous chapter, there would never had adopted the d_{ma} and d_{mo} distinction as we have done in the interpretation of our framework.

On the other hand, we tried to make the case that the developments made in this chapter also illuminate the methodological and philosophical problems posed in this dissertation. For instance, we argued that our framework could help clarify the frontier between behavioral and standard economics on framing effects. The framed descriptive structure with description invariance always characterize standard economics and the normative positions of behavioral economics on framing effects; and the framed descriptive structure with tidy or untidy description-dependence marks different ways of seeing the frontiers depending on whether we are under certainty or uncertainty. Tidy descriptive dependence may be acceptable under certainty to some standard and behavioral economists as rational because preference reversals are prevented; but not under risk, where the preference reversals it can yield are usually deemed irrational by behavioral and standard economists, but needs to be captured in positive models of individual behaviors in behavioral economics (the same reasoning applies to untidy description-dependence, i.e., regardless of whether or not we are under certainty or under risk). Notice however that we have not tackle the issue of theoretical unification regarding framing effects in the three dimensions; this is left for further work.

Likewise, we argued throughout that the d_{ma} and d_{mo} interpertation clarified the positive/normative issue underlying framing effects. This clarification stems from making explicit the axiomatic conditions (i.e., violation of ICD *versus* ICDAC) under which *one* framing effect

can be given *two* different interpretations, i.e., as rational or irrational. It could be argued that this does not clarify anything because 'rationality' is never defined once and for all and then used consistently for evaluative interpretations, i.e., the notion of rationality changes as d_{mo}'s perspective changes. We would respond, following Sen (2002, chap.1), that defining rationality *a priori* and 'once and for all' does not pay full justice to the power of human reason. Being a little more pragmatic about rationality does not imply to take all value judgments of rationality and irrationality as being created equal (i.e., pragmatism is not relativism or nihilism, see Stein 1996). Arguably, the methodological attitude of defining rationality *a priori* and once and for all played a nontrivial role in economists' perspective on description invariance as an 'all or nothing' axiom: either you have it or you don't. By contrast, being a little more pragmatic when it comes to rationality helps not only to see that this is not necessarily the case but also to guide the formal constructions needed to weaken the axiom.

Appendix: Proofs

Proof of proposition 1. Let $x \in X$. By axiom 1, $\approx$ is an equivalence relation, define $[x]_{\approx} := \{a \in X \mid a \approx x\}$ the equivalence class of x by $\approx$. $[x]_{\approx}$ is a subset of X and $\sim$ is an equivalence by axiom 2. Let $[x]_{\approx}/\sim$ be the set of equivalence classes of $[x]_{\approx}$ under $\sim$. Define $\star$ on $[x]_{\approx}/\sim$ as follows: $[a] \star [b] = [a \circ b]$. By axioms 1, if $a, b \in [x]_{\approx}$, then $a \circ b$ is defined and belongs to $[x]_{\approx}$. $\star$ is a closed binary operation on $[x]_{\approx}/\sim$. Observe that $\star$ is well defined since if $a, a', b, b' \in [x]_{\approx}$, $a \sim a'$ and $b \sim b'$, then, by axiom 1, $a \circ b$ and $a' \circ b'$ are defined and by two applications of axiom 6, $a \circ b \sim a' \circ b \sim a' \circ b'$, so $[a] \star [b] = [a'] \star [b']$. We show that $\langle [x]_{\approx}/\sim, \star \rangle$ is a cancellative idempotent commutative semigroup. If $a, b, c \in [x]_{\approx}$, then by axioms 1, $a \circ b$, $b \circ a$, $(a \circ b) \circ c$ and $a \circ (b \circ c)$ are defined. Associativity follows from axiom 4; commutativity follows from axiom 3; idempotence follows from axiom 5; and cancellation follows from axiom 6. It is a classical result that a commutative semigroup can be embedded in a group if and only if it is cancellative. Note that $\langle [x]_{\approx}/\sim, \star \rangle$ is idempotent, it follows that the group extension of $\langle [x]_{\approx}/\sim, \star \rangle$ is a group in which every element is idempotent. Every group has exactly one idempotent element: the identity. It follows that for all $a, b \in [x]_{\approx}$, $a \sim b$. Hence,

$$\forall a, b \in X, \quad a \approx b \Rightarrow a \sim b.$$

Proof of proposition 2. As in proof of Proposition 1, define $[x]_{\approx} := \{a \in X \mid a \approx x\}$ the equivalence class of x by $\approx$; and $\star$ on $[x]_{\approx}/\sim$ as follows: $[a] \star [b] = [a \circ b]$. As axioms 1-5 and 6.1 hold, it is clear that $\star$ is well defined and that $\langle [x]_{\approx}/\sim, \star\rangle$ is an idempotent commutative semigroup. We show that if $a \in X$ and $b \in X$ are generated by the same generators, then $a \sim b$. Let B be an idempotent commutative semigroup generated by $b_1, \ldots, b_n$. For each $x \in B$, the length of x, denoted by $|x|$, is the minimum k such that $x = x_1 x_2 \ldots x_k$, where $x_k \in \{b_1, \ldots, b_n\}$, $k \geq i \geq 1$. Note that $|x| \geq 1$ for all $x \in B$. The largest possible length of an element of any idempotent commutative semigroup with n generators is n. Now, let $a, b \in X$, generated by the same generators. The previous reasoning applies and by axiom 6.1, it is clear that $a \sim b$. Suppose that $[x]_{\approx}$ is finitely generated by n pure descriptions, then $\langle [x]_{\approx}/\sim, \star\rangle$ is finitely generated by n elements. By analogy with the cardinal of the power set of a finite set $\langle [x]_{\approx}/\sim, \star\rangle$ has at most $2^n - 1$ elements.

Proof of proposition 3. In view of proposition 2, we show that axioms 7 and 8 imply that

$$\forall x, y, z \in X, \text{ if } x \approx z \text{ and } x \succsim z, \text{ then } (x \succsim y \succsim z \Rightarrow y \approx x).$$

Assume that $\succsim$ satisfies axioms 7 and 8, take $x, y, z \in X$ such that $x \approx z$, $x \succsim z$, and $x \succsim y \succsim z$), and suppose that not $y \approx x$, hence not $y \approx z$. By axiom 7 there exist $z_1 \approx z$, $x_2 \approx x$, and $y_1, y_2 \approx y$ such that not $z_1 \sim y_1$ and not $x_2 \sim y_2$. Assume $z_1 \succ y_1$ or $y_2 \succ x_2$, applying axiom 8 twice, we obtain $y \succ z$ or $z \succ x$. This contradicts $x \succsim y \succsim z$. Hence, we have $y_1 \succ z_1$ and $x_2 \succ y_2$. By similar reasoning, $x \sim y$, $y \sim z$ and $x \sim z$ are impossible. We can suppose that $x \succ y \succ z$, but again, applying axiom 8 twice, by transitivity we obtain $x \circ z \succ x \circ y$, a contradiction. Therefore, $y \approx x$ (and $y \approx z$).

Proof of proposition 4. Let $\langle X, \approx, \circ\rangle$ be a descriptive structure such that $X/\approx$ is finite and $[x]_{\approx}$ is finitely generated for all $x \in X$. Let $x \in X$, we can explicitly enumerate, in a countable fashion, all possible concatenations on $[x]_{\approx}$. Since this provides a listing of all elements of $[x]_{\approx}$, $[x]_{\approx}$ is countable. $X/\approx$ is finite, whence it is the union of a finite number of countable

sets, hence $X/\approx$ is countable. It is well-known that a countable weakly ordered set admits an utility representation, hence statement (i) is proven. Statement (ii) (respectively (iii), (iv)) is a consequence of (i) and Proposition 1 (respectively Proposition 2, 3).

Proof of proposition 5. Define $r : \mathcal{P}^*(X/\approx) \to \mathcal{P}^*(X)/\approx_{\mathcal{P}}$ by

$$r(I) := \left[\bigcup_{[x]_\approx \in I} [x]_\approx \right]_{\approx_{\mathcal{P}}}.$$

We will show that r is onto. Let $A \in \mathcal{P}^*(X)$. Set $I = \{[x]_\approx \mid x \in A\}$, we will show that $\bigcup_{[x]_\approx \in I}[x]_\approx \approx_{\mathcal{P}} A$, that is $r(I) = [A]_{\approx_{\mathcal{P}}}$. If $x \in A$, by definition, $x \in \bigcup_{[x]_\approx \in I}[x]_\approx$. Conversely, if $y \in \bigcup_{[x]_\approx \in I}[x]_\approx$, there is an $x \in A$ such that $y \approx x$. Therefore $\bigcup_{[x]_\approx \in I}[x]_\approx \approx_{\mathcal{P}} A$. Next we show that r is one-to-one. Suppose that $r(I) = r(J)$ for some $I, J \in \mathcal{P}^*(X/\approx)$. Let $[x]_\approx \in I$, then as $\bigcup_{[x]_\approx \in I}[x]_\approx = \bigcup_{[x]_\approx \in J}[x]_\approx$, and $[x]_\approx \subseteq \bigcup_{[x]_\approx \in I}[x]_\approx$, we have $[x]_\approx \subseteq \bigcup_{[x]_\approx \in J}[x]_\approx$, whence $[x]_\approx \in J$. Hence $I \subseteq J$, by symmetry $I = J$.

Proof of proposition 6. Let X be a finite choice set of described consequences such that $X/\approx$ is not trivial and c an extended choice function. Let $c = c_{\succsim}$ where $\succsim$ is complete and transitive relation on X. To see that c satisfies property γ^+-extended for descriptions, note that if $[x]_\approx \in c([A]_{\approx_{\mathcal{P}}}, f)$, $[y]_\approx \in c([B]_{\approx_{\mathcal{P}}}, g)$, $[x]_\approx \cap f$ is a singleton, and $[y]_\approx \cap g \subseteq f$, then there exists a unique $x' \in f$ such that $x' \approx x$ and $x' \succsim z$ for all $z \in f$ and there exists $y' \in [y]_\approx$ such that $y' \in f$ and $y' \succsim z$ for all $z \in g$. It follows that $x' \succsim z$ for all $z \in f \cup g$. Let $h \in [A]_{\approx_{\mathcal{P}}} \cup [B]_{\approx_{\mathcal{P}}}$ such that $[x]_\approx \cap f \subseteq h \subseteq f \cup g$, whence $x' \in h$, hence $[x]_\approx \in c([A]_{\approx_{\mathcal{P}}} \cup [B]_{\approx_{\mathcal{P}}}, h)$.

In the other direction, let c be an extended choice function satisfying γ^+-extended for descriptions. Define $\succsim_c$ on X as follows: $x \succsim_c y$ iff $x \not\approx y$ and $\exists A \subseteq X$ such that $x, y \in A$, $[x]_\approx \cap A = \{x\}$, $[y]_\approx \cap A = \{y\}$, and $[x]_\approx \in c([A]_{\approx_{\mathcal{P}}}, A)$. If $x \not\approx y$, since c is nonempty by definition, it is straightforward to show that either $x \succsim_c y$ or $y \succsim_c x$. Suppose that $x \succsim_c y$, $y \succsim_c z$ and $x \not\approx z$. By definition, $x \succsim_c y$ means $x \not\approx y$ and $\exists A_{xy} \subseteq X$ such that $x, y \in A_{xy}$, $[x]_\approx \cap A_{xy} = \{x\}$, $[y]_\approx \cap A_{xy} = \{y\}$, and $[x]_\approx \in c([A_{xy}]_{\approx_{\mathcal{P}}}, A_{xy})$. And $y \succsim_c z$ means $y \not\approx z$ and $\exists A_{yz} \subseteq X$ such that $y, z \in A_{yz}$, $[y]_\approx \cap A_{yz} = \{y\}$, $[z]_\approx \cap A_{yz} = \{z\}$, and $[y]_\approx \in c([A_{yz}]_{\approx_{\mathcal{P}}}, A_{yz})$. Let A_{xz} be the maximal subset (w.r.t inclusion) of $A_{xy} \cup A_{yz}$ which belongs to $[A_{xy}]_{\approx_{\mathcal{P}}} \cup [A_{yz}]_{\approx_{\mathcal{P}}}$ and includes x, z as only representative of $[x]_\approx, [z]_\approx$ respectively. Whence, $[x]_\approx \in c([A_{xy}]_{\approx_{\mathcal{P}}}, A_{xy})$, $[y]_\approx \in c([A_{yz}]_{\approx_{\mathcal{P}}}, A_{yz})$,

$[x]_{\approx} \cap A_{xy}$ is a singleton, and $[y]_{\approx} \cap A_{xy} = \{y\} \subseteq A_{xy}$. Hence, γ^+-extended for descriptions applies, as A_{xz} belongs to $[A_{xy}]_{\approx_{\mathcal{P}}} \cup [A_{yz}]_{\approx_{\mathcal{P}}}$ and $[x]_{\approx} \cap A_{xy} = \{x\} \subseteq A_{xz} \subseteq A_{xy} \cup A_{yz}$, it is true that $[x]_{\approx} \in c([A_{xz}]_{\approx_{\mathcal{P}}}, A_{xz})$. Hence $zx \succsim_c z$ because $x \not\approx z$. Now, if $S \subseteq X$ is such that for all $x, y \in S$, $x \not\approx y$ whenever $x \neq y$, then the restriction of $\succsim_c$ is a weak order on S. It is easy to see that $\succsim_c$ is Suzumura consistent, let $\succsim_c^{tr}$ be its transitive closure. If there exits x, y incomparable for $\succsim_c^{tr}$, then necessarily $x \approx y$ and there exist z_1, z_2 such that $z_1, z_2 \not\approx x, y$ and $z_1 \succsim_c^{tr} x, y \succsim_c^{tr} z_2$ (or just z_1 if x, y are bottom elements or just z_2 if x, y are top elements). We can easily complete $\succsim_c^{tr}$ consistently. Let $\succsim_c^{o}$ be such completion. It is straightforward to show that $c = c_{\succsim_c^{o}}$.

Proof of proposition 7. Let X be a set of described consequences and c be a choice function for X. Define $\succsim_c$ on X as follows: $x \succsim_c y$ iff $\exists S \subseteq X$, finite, such that $x, y \in S$, and $x \in c(S)$. Given a binary relation $\succsim$ on X, the induced choice function $c_{\succsim}$ is defined by

$$c_{\succsim}(S) = \{x \in S \mid x \succsim y,\ \forall y \in S\}.$$

It is a classical result that c satisfies $C2$ if, and only if, there exists a binary relation $\succsim$ on X such that $c = c_{\succsim}$ and $\succsim$ satisfies axiom A_2, that is, $\succsim$ is a weak order. We prove the proposition for the conditions $C5$ (Axiom 5) and $C6.1$ (Axiom 6.1). The other cases are similar and/or follow from basic definitions. Let $c = c_{\succsim}$ where $\succsim$ is a binary relation on X that satisfies axiom 5. Suppose that $x \in c(S)$ for some S finite. By definition, $x \succsim y$, $\forall y \in S$, but $x \sim x \circ x$ by hypothesis, whence $x \sim x \circ x \succsim y$, $\forall y \in S$, it follows that $x, x \circ x \in c(S \cup \{x \circ x\})$. Conversely, let c be a choice function satisfying $C5$ and $\succsim_c$ the binary relation induced by c. By definition, $\emptyset \neq c(\{x\}) \subseteq \{x\}$, then clearly $c(\{x\}) = \{x\}$. By $C5$, it is also clear that $x \succsim_c x \circ x$ and $x \circ x \succsim_c x$. Thus $\succsim_c$ satisfies axiom 5. Finally, we prove the proposition for $C6.1$. Let $c = c_{\succsim}$ where $\succsim$ is a binary relation on X that satisfies axiom 6.1. Suppose that $x, y \in c(S)$ and $x \circ z, y \circ z \in T$ for some S, T finite. It is clear that $x \sim y$, whence $x \circ z \sim y \circ z$. It follows that $C6.1$ holds. Conversely, let c be a choice function satisfying $C6.1$ and $\succsim_c$ the binary relation induced by c. Suppose that $x \circ z$ and $y \circ z$ are defined and that $x \sim_c y$. Set $S = \{x, y\}$ and $T = \{x \circ z, y \circ z\}$. Then $C6.1$ applies, $c(T)$ is nonempty by definition, it follows that $x \circ z \sim_c y \circ z$.

General conclusion

This dissertation proposed to scrutinize behavioral economics from a methodological perspective that makes explicit the role of language in economic rationality. As explained in the general introduction, we emphasized this methodological issue because because it had not been raised previously around behavioral economics, by contrast with the other three related methodological issues we also tackled: the issue of interdisciplinarity between economics and *P*sychology, the positive/normative issue within models of individual behaviors, and the problem of theoretical unification in the three dimensions of economic behaviors. Throughout the chapters, we have investigated how these four sets of methodological issues intersect concretely within the scientific literature around behavioral economics.

In chapter 1, we compared the founding work of the father of behavioral economics, Thaler, with the work of Sen whose entanglement thesis on the positive/normative issue we endorsed. This was tantamount to compare two different criticisms of the same models of economic behaviors under certainty. Though both criticisms started from the same observation, namely that the positive and normative dimensions of standard models of individual behaviors need to be articulated differently, they ended in radically different conclusions. Through an interdisciplinary relation with ethics, moral and political philosophy, Sen proposes that the normative dimension can be developed from careful scrutiny of empirical facts; i.e., rationality should not be defined *a priori* and once and for all. Through an interdisciplinary relation with *P*sychology, Thaler proposes that the normative dimension should be separated from the positive one; i.e., standard models provide an adequate notion of rationality that can be used as a benchmark to evaluate empirical behaviors. Two features of individual decision making revealed by the comparison have been investigated further in the other chapters. On the one hand, there was

the communicative structure of choices: decision problems do not fall from the sky, they are posed by decision modelers to decision makers. We suggested Searle's theory of speech acts to flesh out this communicative structure in a way that is relevant for rational choice theory. On the other hand, despite restricting ourselves to the economic analysis of individual behaviors under certainty, the three dimensions of uncertainty, time and other people kept popping up in various ways.

Thus, in chapter 2, we scrutinized the challenges posed by behavioral economics to the standard models used in these three dimensions. We noticed a rupture between, on the one hand, the classical challenges posed by behavioral economics to the three dimensions separately, and, on the other hand, more recent challenges involving interactions across dimensions. It was argued that the marking of linguistic distinctions within the three dimensions (e.g., 'now' and 'later', 'certainly' and 'probably', 'you' and 'me') is a feature of the communicative structure of choices that constitutes the conditions of possibility for both the challenges within and across dimensions. Several contrasts were made about standard and behavioral economists' accounts of behaviors in the three dimensions. On the issue of interdisciplinarity between economics and *P*sychology, the dimensions of uncertainty and time have clear roots in respectively the works of Kahneman and Tversky, and Ainslie. This contrasts with the dimension of other people where no such root in *P*sychology is identifiable. On the positive/normative issue, the three dimensions are connected by the doctrine of consequentialism, from which value judgments of rationality and irrationality are derived. On this point, we saw that, historically, there has been a slow shift in standard decision theory from a primacy of the dimension of uncertainty (over the two others) to a primacy of the dimension of time in these consequentialist value judgments. The chapter ended by emphasizing how neither standard nor behavioral economics' models of individual behaviors could explain the full set of behavioral regularities constitutive of the recent challenges from interactions across dimensions. Accounting for interactions across dimensions was therefore suggested as a potentially fruitful area of convergence for behavioral and standard economics, at least for the purpose of theoretical unification.

This suggestion was developed further in chapter 3 by scrutinizing the dual models that were explicitly seeking theoretical unification and eventually theoretical reconciliation between behavioral and standard economics. There, we studied the role of language in economic ratio-

nality as constructed by economists in their models, abstracting from economic agents' uses of language in their economic behaviors. To do so, we scrutinized the interactions between, on the one hand, the formal and technical languages from economics, and, on the other hand, a new technical language borrowed from the *P*sychology of self-control. Because self-control problems are primarily about the decision maker's relation to time, the historical shift from a primacy of the dimension of uncertainty to the primacy of the dimension of time observed in the previous chapter finds its paradigmatic illustration in dual models. It was argued that only Fudenberg and Levine's dual model tackles the issue of interactions *across* the three dimensions, hence providing a promising avenue of theoretical unification and reconciliation between standard and behavioral economics. Under their account, the standard notion of consistency is modified by conceptualizing the mental limits responsible for self-control problems as an inescapable constraint under which decision makers optimize. Finally, it was shown that Fudenberg and Levine, along with other dual modelers, point at framing phenomena as the main limit to theoretical unification.

Hence we turned our attention to what standard and behavioral economists had to say about 'framing' in chapter 4. It was first remarked that the term is being used to refer to an increasingly heterogeneous set of behavioral phenomena. The only common point of these phenomena is that different presentations of 'the same' decision problem can reveal preference reversals. We emphasized the crucial role played by the theoretical conventions of standard models in establishing relations of equivalence between decision problems. The focus was put on framing effects violating the implicit axiom of description invariance, the axiom of standard models that make them blind to the role of language in economic rationality as displayed by economic agents. Despite some studies of violations of description invariance by behavioral economists, notably in the three dimensions and under certainty, this axiom remains relatively understudied in economics in view of the large body of research on the topic in *P*sychology. We proposed to organize what we considered as the key features, from an economist's perspective, of three decades of research by psychologists on violations of description invariance. Hence we abstracted from the role of language in economic rationality as constructed by economists in their models to focus only on the role of language in economic rationality as displayed by economic agents in their behaviors. We emphasized on the conditions under which Tversky and

Kahneman's original violation of description invariance (in the Asian Disease problem), on the one hand, holds empirically, and, on the other hand, can be defended as rational. Among these conditions, we argued that three of them could be especially useful for the purpose of a formal account of framing effects in economics. The first one was about further redescriptions of the same consequences (notably through the use of negation). The second one was about the direct dependence of preference on description (notably justified by experienced utility). And the third one was about the exchanges of tacit information through the uses of ordinary language in the communicative structure of choices (as displayed in Sher and McKenzie's work).

The three underlying sets of results from *P*sychology were then used to motivate, in chapter 5, an axiomatic framework constructed with Dino Borie to deal with violations of description invariance. Our strategy was first to make the axiom explicit in a way that was fine-grained enough to then weaken it, instead of characterizing it as an 'all or nothing' axiom, i.e., either you have it or you don't. We tried to make the case for discriminating different degrees of dependence to descriptions, two of which we characterized axiomatically. It was then shown that under the weaker degree of dependence to descriptions, most framing effects violating description invariance could be represented in standard models, with the traditional equivalence among choice, preference and utility. Throughout, we emphasized an interpretation of our framework in terms of *two* simultaneous but possibly distinct perspectives on *one* decision problem, namely the decision maker's and the decision modeler's. This was intended to place the communicative structure of choices in the foreground. In turn, this allowed to tackle the positive/normative issue of whether violations of description invariance are necessarily irrational by clarifying under which interpretations some conditions rather than others are being violated in a given framing effect. These minor modifications to economists' formal and technical languages were intended as a first step to take into account economically relevant uses of everyday language by economic agents.

The intended contribution of this dissertation was to show how a methodological perspective on the twofold role of language in economic rationality can clarify three main issues (and their connections) underlying the behavioral *versus* standard economics debates: the issue of the theoretical unification regarding the three dimensions of economic rationality, the issue of interdisciplinarity between economics and *P*sychology and the positive/normative issue within

models of individual behaviors. We discussed several types of markers of the frontier between standard and behavioral economics. Theoretical markers are constituted by a series of *p*-&-*P*sychological notions such as loss aversion under certainty and risk, impulsivity over time, various non-self-centered motives regarding other people. Empirical markers are essentially constituted by framing effects. But there are not really any kind of normative markers, as the doctrine of consequentialism is largely accepted on both sides to derive value judgments of irrationality about what lies in the behavioral side of the theoretical and empirical markers. An implications of this is that the behavioral *versus* standard economics debates are blind to Sen's notion of commitment. We have argued that one weakness of this position is that the notion of 'a consequence' shared by most economists does not seem to be well-defined for the purpose of an empirical science.

Furthermore, this dissertation sought to go beyond mere clarification regarding the positive/normative issue and the role of language in the behaviors of economic agents. My intention was to provide a constructive criticism of contributions from behavioral as well as standard economists on both of these points. Following the entanglement thesis, it has been argued that both standard and behavioral economists propose an unsatisfying articulation between the positive and normative dimensions of models of individual behaviors; and that recognizing the entanglement of facts, values and convention can actually be theoretically and empirically fruitful. Furthermore, it has been argued that paying some attention to the role of language in the behaviors of economic agents may *sometimes* show that a seemingly irrational behavior can in fact be defended as rational. The formal contribution on this point was partly intended to show that reflections from, on the one hand, methodology and philosophy, and, on the other hand, standard economic theory, can have a mutually beneficial influence.

In terms of further work, some projects related to this dissertation are already in process. The arguments sketched in chapter 2 about the possibility of critical reinterpretations of existing empirical studies that have used only one dimension in their original interpretations of the data are explored in a series of case studies with Judith Favereau and Cléo Chassonnery-Zaïgouche. The arguments also sketched in chapter 2 about the relations between narrativity and identity regarding dynamic consistency are the object of a paper with Tom Juille. Some of the arguments sketched in both chapter 1 and 2 about the experimental translation of rationality are developed

from an epistemological and historical perspectives in a working paper; the epistemological perspective tries both to ground the experimental translation of rationality in the epistemological theory of reflective equilibrium while at the same time providing a possible empirical operationalization of that theory; the historical perspective discusses the influence (due to the rise of behavioral economics) of economics on *P*sychology in the so-called 'rationality debate'. In the making of the working paper underlying chapter 5, we (together with Dino Borie) conducted some classroom experiments on some implications of our framework; we are discussing the possibility of doing 'real' experiments with some experimental economists from the University of Nice Sophia-Antipolis, notably Ismaël Raifaï with whom we have also discussed the possibility of building a model of framing effects from our axiomatic framework.

I would like to conclude this dissertation with Fritz Machlup. In a paper entitled "If Matter Could Talk", Machlup introduces the methodological issue of the distinction between the social and the natural sciences with a story, part of which goes as follows:

> "As he spoke about random walk of molecules and about molecular collisions at various pressures, someone shouted, "Stop that nonsense!" When he looked around to see which student had made this impertinent remark, the voice continued. It was obviously coming from the protective chamber with the suspended mirror, whose movements were being tracked by the fluctuations of a reflected light beam. This is what he heard: "It is time that you cease and desist from misleading your students. What you teach about us molecules is simply not true. This is no random walk and we are not pushing one another all over the place. We know where we are going and why. If you will listen, we shall be glad to tell you." He had not waited for more, but had rushed here to report and get Professor R. to witness the event and to hear what the molecules were about to tell.
> "Oh," said Professor R., "you mean they are going to tell us what they *think* they are doing. By all means, let them go ahead."" (Machlup 1978, p.310-311)

The goal of Machlup (1978 [1969]) is to argue that one methodological specificity of the social sciences is that their "data and problems" (ibid, p.319) are soaked with language uses, not only coming from the scientists (this is not a specificity of the social sciences), but also from the objects of study. More precisely, he wants to argue that the social sciences are specific because of the possibility of "contradictory communication" (ibid) between the scientists and their subjects. While a sandstone cannot say to a geologist 'No way, we are *not* from this family you call 'Sedimentary Rock', we are very different from these individuals', a decision maker can say to a decision theorist 'No way, I am *not* as you put it 'irrational' or 'quasi-rational', I have

plenty of reasons to maintain my preferences as they are, and I do not 'reverse' them'. As we have seen, this is roughly what happened around the Allais paradoxes for a brief historical moment by the end of the 1970s. Decision scientists (MacCrimmon, Larson, Slovic, Tversky among others) *listened* to their subjects instead of just trying to merely observe their behaviors; and this contributed to nontrivial theoretical developments. As we have also seen, this is not without methodological dangers, notably because there are many subtle features of a *conversation* that cannot be controlled in the way the minimalist *communication* in experiments can be controlled.

In any case, the uses of language in economic rationality as constructed by economists and as displayed by economic agents, notably within the communicative structure of choices, were not restricted to the issue of contradictory communication. It is easier to state the issues we were interested in by turning Machlup's question upside-down: what if economic agents couldn't talk? If that were the case, then, as argued in chapter 2, none of the classical or more recent challenges posed by behavioral economics in the three dimensions of economic rationality could have been posed in the first place; as argued in chapter 4, framing effects would not exist; and if we keep in mind that economists are economic agents, then, as argued in chapter 1 and chapter 3, the construction of models of individual behaviors would be severely impaired, if not totally impossible. One warning I did receive about my focus on the role of language was 'Don't go there, once you are interested in language, you see language everywhere'. Not only I agree, but I agree that there are underlying methodological dangers to that. However, language may not be strictly speaking *everywhere*, but it is quite pervasive in ways that are not trivial for economics. In my opinion, there are just as great methodological dangers in *denying* the pervasiveness of a pervasive phenomenon.

Bibliography

Abdellaoui, M., Barrios, C. and Wakker, P. P.: 2007, Reconciling Introspective Utility With Revealed Preference: Experimental Arguments Based on Prospect Theory, *Journal of Econometrics* **138**(1), 356–378.

Abdellaoui, M., Bleichrodt, H., L'Haridon, O. and Paraschiv, C.: 2013, Is There One Unifying Concept of Utility?An Experimental Comparison of Utility Under Risk and Utility Over Time, *Management Science* **59**(9), 2153–2169.

Ahn, D. S. and Ergin, H.: 2010, Framing Contingencies, *Econometrica* **78**(2), 655–695.

Ainslie, G.: 1975, Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control, *Psychological Bulletin* **82**(4), 463–496.

Ainslie, G.: 1992, *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*, Cambridge University Press, Cambridge, England.

Ainslie, G.: 2001, *Breakdown of Will*, Cambridge University Press, Cambridge, England.

Ainslie, G.: 2012, Pure Hyperbolic Discount Curves Predict "eyes open" Self-control, *Theory and Decision* **73**(1), 3–34.

Alcouffe, A.: 2013, Economie des langues et des politiques linguistiques, *in* G. Kremnitz (ed.), *Histoire sociale des langues de France*, Presse Universitaire de Rennes, Rennes, pp. 209–224.

Allais, M.: 1953, Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'école americaine, *Econometrica* **21**(4), 503–546.

Allais, M.: 1979a, The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School, *in* M. Allais and O. Hagen (eds), *Expected Utility Hypothesis and the Allais Paradox*, Springer, Dordrecht, pp. 27–145.

Allais, M.: 1979b, The So-Called Allais Paradix and Ratrional Decisions Under Uncertainty, *in* M. Allais and O. Hagen (eds), *Expected Utility Hypothesis and the Allais Paradox*, Springer, Dordrecht, pp. 437–682.

Alós-ferrer, C. and Strack, F.: 2014, From Dual Processes to Multiple Selves : Implications for Economic Behavior, *Journal of Economic Psychology* **41**(April), 1–11.

Amadae, S. M.: 2003, *Rationalizing Capitalist Democracy: The Cold War Origin of Rational Choice Liberalism*, The University of Chicago Press, Chicago and London.

Andersen, S., Harrison, G. W., Lau, M. I. and Rutström, E. E.: 2008, Eliciting Risk and Time Preferences, *Econometrica* **76**(3), 583–618.

Andersen, S., Harrison, G. W., Lau, M. I. and Rutström, E. E.: 2014a, Discounting Behavior: A Reconsideration, *European Economic Review* **71**(Octobre), 15–33.

Andersen, S., Harrison, G. W., Lau, M. I. and Rutström, E. E.: 2014b, Dual Criteria Decision, *Journal of Economic Psychology* **41**(April), 101–113.

Andreoni, J.: 1995, Warm-glow versus Cold-prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments, *The Quarterly Journal of Economics* **110**(1), 1–21.

Andreoni, J., Aydin, D., Barton, B., Bernheim, D. B. and Naecker, J.: 2016, When Fair Isn't Fair: Sophisticated Time Inconsistency in Social Preferences, *Working Paper* .
URL: *http://web.stanford.edu/ daydin/DAydin_Fair.pdf*

Andreoni, J., Kuhn, M. A. and Sprenger, C.: 2015, Measuring Time Preferences: A Comparison of Experimental Methods, *Journal of Economic Behavior & Organization* **116**(August), 451–464.

Andreoni, J. and Miller, J.: 2002, Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism, *Econometrica* **70**(2), 737–753.

Andreoni, J. and Sprenger, C.: 2012, Risk Preferences Are Not Time Preferences, *American Economic Review* **102**(7), 3357–3376.
URL: *http://pubs.aeaweb.org/doi/abs/10.1257/aer.102.7.3357*

Andreoni, J. and Sprenger, C.: 2015, Risk Preferences Are Not Time Preferences: Reply, *The American Economic Review* **105**(7), 2287–2293.

Angner, E. and Loewenstein, G.: 2012, Behavioral Economics, *in* U. Mäki (ed.), *Handbook of the Philosophy of Science, Vol.5: Philosophy of Economics*, Elsevier, Oxford, pp. 641–690.

Anscombe, F. J. and Aumann, R. J.: 1963, A Definition of Subjective Probability, *The Annals of Mathematical Statistics* **34**(1), 199–205.

Anscombe, G. E. M.: 1969, Causality and Extensionality, *The Journal of Philosophy* **66**(6), 152–153.

Antinyan, A.: 2014, Loss and Other-Regarding Preferences, *Working Paper* .
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2413022*

Arena, R.: 2012, Economic Rationality and the Emergence of Institutions: a Schumpeterian View, *in* H. M. Krämer, H. D. Kurz and H.-M. Trautwein (eds), *Macroeconomics and the history of economic thought: Festschrift in honour of Harald Hagemann*, Routledge, Oxford, pp. 329–337.

Armatte, M.: 2004, L'axiomatisation et les théories économiques: un commentaire, *Revue Economique* **55**(1), 130–142.

Arrow, K. J.: 1982, Risk Perception in Psychology and Economics, *Economic Inquiry* **XX**, 1–9.

Ashraf, N., Camerer, C. F. and Loewenstein, G.: 2005, Adam Smith, Behavioral Economist, *Journal of Economic Perspectives* **19**(3), 131–145.

Atkinson, G., Dietz, S., Helgeson, J., Hepburn, C. and Sælen, H.: 2009, Siblings, Not Triplets: Social Preferences for Risk, Inequality and Time in Discounting Climate Change, *Economics: The Open-Access, Open-Assessment E-Journal* **3**, 0–29.
URL: *http://www.economics-ejournal.org/economics/journalarticles/2009-26*

Attardo, S.: 1997, Locutionary and Perlocutionary Cooperation: The Perlocutionary Cooperative Principle, *Journal of Pragmatics* **27**(6), 753–779.

Aumann, R. J.: 1987, Letter to Leonard Savage, 8 January 1971, *in* J. H. Dreze (ed.), *Essays on Economic Decisions under Uncertainty*, Cambridge University Press, Cambridge, England, pp. 76–78.

Aumann, R. J. and Dreze, J. H.: 2009, Assessing Strategic Risk, *American Economic Journal: Microeconomics* **1**(1), 1–16.

Austin, J. L.: 1975, *How to Do Things with Words*, Harvard University Press, Cambridge, MA.

Baccelli, J.: 2013a, Le comportement et le concept de choix, *Dialogue2* **52**(1), 43–60.

Baccelli, J.: 2013b, Re-Specifying Options: A Solution to teh Paradoxes of Preference Theory, *Working Paper* .

Baccelli, J.: 2016, L'analyse axiomatique et l'attitude par rapport au risque, *Revue Economique* (Forthcoming).

Bach, K.: 2006, Pragmatics and the Philosophy of Language, *in* L. R. Horn and G. Ward (eds), *The Handbook of Pragmatics*, Blackwell, Malden, pp. 463–487.

Bacharach, M. O.: 1990, Commodities, Language, and Desire, *The Journal of Philosophy* **87**(7), 346–368.

Bacharach, M. O.: 1994, The Epistemic Structure of a Theory of a Game, *Theory and Decision* **37**(1), 7–48.

Bacharach, M. O.: 2003, Framing and Cognition in Economics: The Bad News and the Good, *in* N. Dimitri, M. Basili and I. Gilboa (eds), *Cognitive Processes and Economic Behaviour*, Routledge, London, pp. 63–74.

Backhouse, R. E.: 1998, On Knowing One's Place: The Role of Formalism in Economics, *The Economic Journal* **108**(451), 1859–1869.

Barberis, N. C.: 2013, Thirty Years of Prospect Theory in Economics: A Review and Assessment, *Journal of Economic Perspectives* **27**(1), 173–196.

Barberis, N., Huang, M. and Thaler, R. H.: 2006, Individual Preferences, Monetary Gambles, and Stock Market Participation: A Case for Narrow Framing, *The American Economic Review* **96**(4), 1069–1090.

Bardsley, N.: 2005, Experimental Economics and the Artificiality of Alteration, *Journal of Economic Methodology* **12**(2), 239–251.

Bardsley, N.: 2008, Dictator Game Giving: altruism or artefact?, *Experimental Economics* **11**(2), 122–133.

Bardsley, N., Cubitt, R. P., Loomes, G., Moffatt, P., Starmer, C. and Sugden, R.: 2010, *Experimental Economics: Rethinking the Rules*, Princeton University Press, Princeton.

Bargh, J. A.: 1999, The Cognitive Monster: The Case Against the Controllability of Automatic Stereotype Effects, *in* S. Chaiken and Y. Trope (eds), *Dual-process Theories in Social Psychology*, Guilford Press, New York, pp. 361–382.

Baron, J.: 1994, Nonconsequentialist Decisions, *Behavioral and Brain Sciences* **17**(1), 1–42.

Baucells, M. and Heukamp, F. H.: 2010, Common Ratio Using Delay, *Theory and Decision* **68**(1-2), 149–158.

Baucells, M. and Heukamp, F. H.: 2012, Probability and Time Trade-off, *Management Science* **58**(4), 831–842.

Baujard, A.: 2016, Welfare Economics, *in* G. Faccarello and H. D. Kurz (eds), *Handbook of the History of Economic Analysis, Vol.1*, forthcomin edn, Edward Elgar Publishing, Cheltenham.

Baumeister, R. F.: 2003, The Psychology of Irrationality: Why People Make Foolish, Self-Defeating Choices, *in* I. Brocas and J. D. Carrillo (eds), *The Psychology of Economic Decisions (Volume 1: Rationality and Well-Being)*, Oxford University Press, Oxford, pp. 3–16.

Baumeister, R. F.: 2005, Rethinking Self-Esteem: Why Nonprofits Should Stop Pushing Self-esteem and Start Endorsing Self-control, *Stanford Social Innovation Review* (Winter), 34–41.

Baumeister, R. F.: 2015, Self-Control: The Secret to Life's Successes, *Scientific American* **312**(4).

Bell, D., Raiffa, H. and Tversky, A.: 1988, *Decision Making: Descriptive, Normative, and Prescriptive Interactions*, Cambridge University Press, Cambridge, England.

Bénabou, R. and Tirole, J.: 2011, Identity, Morals, and Taboos: Beliefs as Assets, *The Quarterly Journal of Economics* **126**(2), 805–855.

Benartzi, S., Peleg, E. and Thaler, R. H.: 2013, Choice Architecture and Retirement Saving Plans, *in* E. Shafir (ed.), *The Behavioral Foundations of Public Policy*, Princeton University Press, Princeton and Oxford, pp. 245–263.

Benhabib, J. and Bisin, A.: 2005, Modeling Internal Commitment Mechanisms and Self-control: A Neuroeconomics Approach to Consumption-saving Decisions, *Games and Economic Behavior* **52**(2), 460–492.

Benhabib, J. and Bisin, A.: 2008, Choice and Process: Theory Ahead of Measurement, *in* A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics. A Handbook*, Oxford University Press, New York, pp. 320–335.

Benhabib, J., Bisin, A. and Schotter, A.: 2010, Present-bias, Quasi-hyperbolic Discounting, and Fixed Costs, *Games and Economic Behavior* **69**(2), 205–223.

Bernheim, D. B.: 2009, On the Potential of Neuroeconomics : A Critical (but Hopeful) Appraisal, *American Economic Journal: Microeconomics* **1**(2), 1–41.

Bernheim, D. B. and Rangel, A.: 2004, Addiction and Cue-Triggered Decision Processes, *The American Economic Review* **94**(5), 1558–1590.

Bernheim, D. B. and Rangel, A.: 2007a, Behavioral Public Economics : Welfare and Policy Analysis with Non-standard Decision-Makers, *in* P. Diamond and H. Vartianinen (eds), *Behavioral Economics and its Applications*, Princeton University Press, Princeton and Oxford, pp. 7–77.

Bernheim, D. B. and Rangel, A.: 2007b, Toward Choice-Theoretic Foundations for Behavioral Welfare Economics, *The American Economic Review* **97**(2), 464–470.

Bernheim, D. B. and Rangel, A.: 2008, Choice-theoretic Foundations for Behavioral Welfare Economics, *in* A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics: A Handbook*, Oxford University Press, New York, pp. 155–192.

Bernheim, D. B. and Rangel, A.: 2009, Beyond Revealed Preference: Choice-theoretic Foundations for Behavioral Welfare Economics, *Quarterly Journal of Economics* **124**(1), 51–104.

Berridge, K. C.: 2003, Irrational Pursuits: Hyper-Incentives From a Visceral Brain, *in* I. Brocas and J. D. Carrillo (eds), *The Psychology of Economic Decisions (Volume 1: Rationality and Well-Being)*, Oxford University Press, Oxford, pp. 17–40.

Berridge, K. C. and O'Doherty, J. P.: 2014, From Experienced Utility to Decision Utility, *in* P. W. Glimcher and E. Fehr (eds), *Handbook of Neuroeconomics (2nd ed.)*, Elsevier, New York, pp. 335–351.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E. and Zinman, J.: 2010, What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment, *The Quarterly Journal of Economics* **125**(1), 263–306.

Binmore, K. and Shaked, A.: 2010a, Experimental Economics: Where Next?, *Journal of Economic Behavior & Organization* **73**(1), 87–100.

Binmore, K. and Shaked, A.: 2010b, Experimental Economics: Where Next? Rejoinder, *Journal of Economic Behavior & Organization* **73**(1), 120–121.

Birnbaum, M. H.: 2005, A Comparison of Five Models that Predict Violations of First-Order Stochastic Dominance in Risky Decision Making, *The Journal of Risk and Uncertainty* **31**(3), 263–287.

Blaug, M.: 1992, *The Methodology of Positive Economics – Or How Economists Explain (Second Edition)*, Cambridge University Press, Cambridge, England.

Blaug, M.: 1998, The Positive-normative Distinction, *in* J. B. Davis, W. Hands and U. Mäki (eds), *Handbook of Economic Methodology*, Elgar, Cheltenham, pp. 370–4.

Bleichrodt, H., Li, C., Moscati, I. and Wakker, P. P.: 2016, Nash Was a First to Axiomatize Expected Utility, *Theory and Decision* (Forthcoming).

Bleichrodt, H., Rohde, K. I. and Wakker, P. P.: 2008, Koopman's Constant Discounting for Intertemporal Choice: A Simplification and a Generalization, *Journal of Mathematical Psychology* **52**(6), 341–347.

Bless, H., Betsch, T. and Franzen, A.: 1998, Framing the Framing Effect: The Impact of Context Cues on Solutions to the 'Asian Disease' Problem, *European Journal of Social Psychology* **28**(2), 287–291.

Blinder, A. S., Ehrmann, M., Fratzscher, M., de Haan, J. and Jansen, D.-J.: 2008, Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence, *Journal of Economic Literature* **46**(4), 910–945.

Bohnet, I. and Frey, B. S.: 1999, Social Distance and Other-regarding Behavior in Dictator Games: Comment, *The American Economic Review* **89**(1), 335–339.

Bolton, G. E. and Ockenfels, A.: 2000, ERC: A Theory of Equity, Reciprocity, and Competition, *The American Economic Review* **90**(1), 166–193.

Bolton, G. E. and Ockenfels, A.: 2010, Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States: Comment, *The American Economic Review* **100**(1), 628–633.

Bonnefon, J.-F.: 2013, New Ambitions for a New Paradigm: Putting the Psychology of Reasoning at the Service of Humanity, *Thinking & Reasoning* **19**(3), 381–398.

Börgers, T.: 2015, *An Introduction to the Theory of Mechanism Design*, Oxford University Press, New York.

Bossert, W. and Suzumura, K.: 2010, *Consistency, Choice, and Rationality*, Harvard University Press, Cambridge, MA.

Boumans, M.: 2005, *How Economists Model the World into Numbers*, Routledge, London and New-york.

Boumans, M.: 2007, Invariance and Calibration, *in* M. Boumans (ed.), *Measurement in Economics. A Handbook*, Elsevier Inc., London, chapter 9, pp. 231–247.

Bourgeois-Gironde, S. and Giraud, R.: 2009, Framing Effects as Violations of Extensionality, *Theory and Decision* **67**(4), 385–404.

Bovens, L.: 2015, Evaluating Risky Prospects: the Distribution View, *Analysis* **75**(2), 243–253.

Bradley, R. and Stefansson, H. O.: 2016, Counterfactual Desirability, *British Journal for the Philosophy of Science* (Forthcoming).

Brañas-Garza, P.: 2006, Poverty in Dictator Games: Awekening Solidarity, *Journal of Economic Behavior & Organization* **60**(3), 306–320.

Brocas, I.: 2012, Information Processing and Decision-making: Evidence From the Brain Sciences and Implications for Economics, *Journal of Economic Behavior & Organization* **83**(3), 292–310.

Brocas, I. and Carrillo, J. D.: 2003a, Information and Self-Control, *in* I. Brocas and J. D. Carrillo (eds), *The Psychology of Economic Decisions (Volume 1: Rationality and Well-Being)*, Oxford University Press, Oxford, pp. 89–104.

Brocas, I. and Carrillo, J. D.: 2003b, Introduction, *in* I. Brocas and J. D. Carrillo (eds), *The Psychology of Economic Decisions (Volume 1: Rationality and Well-Being)*, Oxford University Press, Oxford, pp. xiii–xxxii.

Brocas, I. and Carrillo, J. D.: 2003c, *The Psychology of Economic Decisions (Volume1: Rationality and Well-Being)*, Oxford University Press, Oxford.

Brocas, I. and Carrillo, J. D.: 2004a, Introduction, *in* I. Brocas and J. D. Carrillo (eds), *The Psychology of Economic Decisions (Volume 2: Reasons and Choices)*, Oxford University Press, Oxford, pp. xv–xxvi.

Brocas, I. and Carrillo, J. D.: 2004b, *The Psychology of Economic Decisions (Volume 2: Reasons and Choices)*, Oxford University Press, Oxford.

Brocas, I. and Carrillo, J. D.: 2008, The Brain as a Hierarchical Organization, *The American Economic Review* **98**(4), 1312–1346.

Brocas, I. and Carrillo, J. D.: 2014, Dual-process Theories of Decision-making: A Selective Survey, *Journal of Economic Psychology* **41**(April), 45–54.

Brocas, I., Carrillo, J. D. and Dewatripont, M.: 2004, Commitment Devices under Self-control Problems: An Overview, *in* I. Brocas and J. D. Carrillo (eds), *The Psychology of Economic Decisions (Volume 2: Reasons and Choices)*, Oxford University Press, Oxford, pp. 49–66.

Brochier, H.: 1995, L'économie comme science positive et normative, *in* G. Duménil and D. Lévy (eds), *L'économie devient-elle une science dure?*, Economica, Paris, pp. 38–54.

Brock, M. J., Lange, A. and Ozbay, E. Y.: 2013, Dictating the Risk: Experimental Evidence on Giving in Risky Environments, *The American Economic Review* **103**(1), 415–437.

Broome, J.: 1991, *Weighing Goods: Equality, Uncertainty and Time*, Basil Blackwell, Oxford.

Bruno, B.: 2013, Reconciling Economics and Psychology on Intrinsic Motivation, *Journal of Neuroscience, Psychology, and Economics* **6**(2), 136–149.

Bughart, D., Epper, T. and Fehr, E.: 2015, The Two Faces of Independence: Betweenness and Homotheticity, *Working Paper* .
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2523486*

Cacioppo, J. T. and Petty, R.: 1989, Effects of Message Repetition on Argument Processing, Recall, and Persuasion, *Basic and Applied Social Psychology* **10**(1), 3–12.

Camerer, C. F.: 1995, Individual Decision Making, *in* J. H. Kagel and A. E. Roth (eds), *The Handbook of Experimental Economics*, Princeton University Press, Princeton, pp. 587–703.

Camerer, C. F.: 1999, Behavioral Economics: Reunifying Psychology and Economics, *Proceedings of the National Academy of Sciences* **96**(19), 10575–10577.

Camerer, C. F.: 2003, *Behavioral Game Theoy: Experiments in Strategic Interaction*, Princeton University Press, Princeton.

Camerer, C. F.: 2004, Behavioral Game Theory: Predicting Human Behavior in Strategic Situations, *in* C. F. Camerer, G. Loewenstein and M. Rabin (eds), *Advances in Behavioral Economics*, Princeton University Press, Princeton, pp. 374–390.

Camerer, C. F.: 2006, Behavioral Economics, *in* R. Blundell, W. Newey and T. Persson (eds), *World Congress of the Econometric Society*, Cambridge University Press, London, pp. 181–214.

Camerer, C. F.: 2008, The Case for Mindful Economics, *in* A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics*, Oxford University Press, New York, pp. 43–69.

Camerer, C. F.: 2015, The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List, *in* G. Fréchette and A. Schotter (eds), *Handbook of Experimental Economic Methodology*, Oxford University Press, New York, pp. 249–295.

Camerer, C. F., Issacharoff, S., Loewenstein, G., O'Donoghue, T. and Rabin, M.: 2003, Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism', *University of Pensylvania Law Review* **151**(3), 1211–1254.

Camerer, C. F. and Loewenstein, G.: 2004, Behavioral Economics: Past, Present, Future, *in* C. F. Camerer, G. F. Loewenstein and M. Rabin (eds), *Advances in Behavioral Economics*, Princeton University Press, Princeton, pp. 3–51.

Camerer, C. F., Loewenstein, G. F. and Rabin, M.: 2004, *Advances in Behavioral Economics*, Princeton University Press, Princeton.

Camerer, C. F., Loewenstein, G. and Prelec, D.: 2005, Neuroeconomics: How Neuroscience Can Inform Economics, *Journal of Economic Literature* **XLIII**(1), 9–64.

Caplin, A. and Schotter, A.: 2008a, *The Foundations of Positive and Normative Economics : A Handbook*, Oxford University Press, New York.

Caplin, A. and Schotter, A.: 2008b, Volume Introduction, *in* A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics: A Handbook*, Oxford University Press, New York, pp. xv–xxii.

Cat, J.: 2013, The Unity of Science, *Stanford Encyclopedia of Philosophy* .
URL: *http://plato.stanford.edu/entries/scientific-unity/*

Chao, H.-K.: 2007, Structure, *in* M. Boumans (ed.), *Measurement in Economics. A Handbook*, Elsevier Inc., London, chapter 11, pp. 271–294.

Charness, G. and Source, M. R.: 2002, Understanding Social Preferences with Simple Tests, *The Quarterly Journal of Economics* **117**(3), 817–869.

Chateauneuf, A. and Wakker, P. P.: 1999, An Axiomatization of Cumulative Prospect Theory for Decision Under Risk, *Journal of Risk and Uncertainty* **18**(2), 137–145.

Chen, K. M.: 2013, The Effect of Language on Economic Behavior : Evidence from Savings Rates , Health Behaviors , and Retirement Assets, *The American Economic Review* **130**(2), 690–731.

Chick, C. F., Reyna, V. F. and Corbin, J. C.: 2015, Framing Effects Are Robust to Linguistic Disambiguation: A Critical Test of Contemporary Theory, *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Advance Online Publication).
URL: *http://dx.doi.org/10.1037/xlm0000158*

Chipman, J. S., Hurwicz, L., Richter, M. K. and Sonnenschein Hugo, F.: 1971, *Preferences, Utility and Demand : A Minnesota Symposium*, Hartcourt Brace Jovanovich, New York.

Chisholm, R. M.: 1941, Sextus Empiricus and Modern Empiricism, *Philosophy of Science1* **8**(3), 371–384.

Chuang, Y. and Schechter, L.: 2015, Stability of Experimental and Survey Measures of Risk, Time, and Social Preferences: A Review and Some New Results, *Journal of Development Economics* **117**(November), 151–170.

Coble, K. H. and Lusk, J. L.: 2010, At the Nexus of Risk and Time Preferences: An Experimental Investigation, *Journal of Risk and Uncertainty* **41**(1), 67–79.

Cohen, J. L.: 1981, Can Human Rationality be Experimentally Demonstrated?, *The Behavioral and brain sciences* **3**(4), 317–370.

Colander, D.: 2000, The Death of Neoclassical Economics, *Journal of the History of Economic Thought* **22**(02), 127–143.

Cot, A. L. and Ferey, S.: 2016, La construction de "faits" économiques d'un nouveau type: éléments d'histoire de l'économie expérimentale, *Working Paper* .
URL: *http://expertise.hec.ca/actualiteeconomique/wp-content/uploads/2015/10/92_1_2_AE_01_Ferey.pdf*

Cowen, T.: 1993, The Scope and Limits of Preference Sovereignty, *Economics and Philosophy* **9**(02), 253–269.

Cozic, M.: 2009, Du concept de conséquence logique: présentation, *in* D. Bonnay and M. Cozic (eds), *Textes clés de philosophie de la logique*, Vrin, Paris, pp. 75–82.

Cubitt, R. P. and Sugden, R.: 2003, Common Knowledge, Salience and Convention: a Reconstruction of David Lewis' Game Theory, *Economics and Philosophy* (2), 175–210.

Danziger, K.: 1997, *Naming the Mind: How Psychology Found its Language*, Sage Publications, London.

Dasgupta, P.: 2005, What Do Economists Analyze and Why: Values or Facts?, *Economics and Philosophy* **21**(02), 221–278.

Dasgupta, P.: 2007, Reply to Putnam and Walsh, *Economics and Philosophy* **23**(03), 365–372.

Dasgupta, P.: 2009, Facts and Values in Modern Economics, *in* D. Ross and H. Kincaid (eds), *The Oxford Handbook of Philosophy of Economics*, Oxford University Press, Oxford, pp. 580–639.

Davis, J. B.: 2006, The Turn in Economics: Neoclassical Dominance to Mainstream Pluralism?, *Journal of Institutional Economics* **2**(1), 1–20.

Davis, J. B.: 2007a, Identity and Commitment: Sen's Fourth Aspect of the Self, *in* F. Peter and H. B. Schmid (eds), *Rationality and Commitment*, Oxford University Press, Oxford, pp. 313–336.

Davis, J. B.: 2007b, The Turn in Economics and the Turn in Economic Methodology, *Journal of Economic Methodology* **14**(3), 275–290.

Davis, J. B.: 2009a, Identity and Individual Economic Agents: A Narrative Approach, *Review of Social Economy* **67**(1), 71–94.

Davis, J. B.: 2009b, The Turn in Recent Economics and Return of Orthodoxy, *Cambridge Journal of Economics* **32**(3), 349–366.

Davis, J. B.: 2011, *Individuals and identity in economics*, Cambridge University Press, New York.

Davis, S.: 1980, Perlocutions, *in* J. R. Searle, F. Kieffer and M. Bierwisch (eds), *Speech Act Theory and Pragmatics*, Reidel, Dordrecht, pp. 37–56.

Davis, W.: 2014, Implicature, *Stanford Encyclopedia of Philosophy* .
URL: *http://stanford.library.usyd.edu.au/entries/implicature/*

De Martino, B., Kumaran, D., Seymour, B. and Dolan, R. J.: 2006, Frames, Biases, and Rational Decision-Making in the Human Brain, *Science* **313**(5787), 684–7.

de Saussure, F.: 1916, *Cours de Linguistique Générale*, Payot & Rivages, Paris.

Dean, M. and Ortoleva, P.: 2015, Is it All Connected? A Testing Ground for Unified Theories of Behavioral Economics Phenomena, *Working Paper* .
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2643355*

Deaton, A. and Muellbauer, J.: 1980, *Economics and Consumer Behavior*, Cambridge University Press, New York.

Debreu, G.: 1954, Representation of a Preference Ordering by a Numerical Function, *in* R. M. Thrall, C. H. Coombs and R. L. Davis (eds), *Decision Processes*, John Wiley & Sons, Inc., Oxford, pp. pp.159–165.

Debreu, G.: 1960, Topological Methods in Cardinal Utility Theory, *in* K. J. Arrow, S. Karlin and P. Suppes (eds), *Mathematical Methods in the Social Sciences*, Stanford University Press, Stanford, pp. 16–26.

Dekel, E. and Lipman, B. L.: 2010, How (Not) to Do Decision Theory, *Annual Review of Economics* **2**(2), 257–82.

del Corral, M. and Bonilla, J. Z.: 2008, Symposium on Language and Games. Introduction: Also Sprach der homo oeconomicus, *Journal of Economic Methodology* **15**(3), 241–244.

DellaVigna, S.: 2009, Psychology and Economics: Evidence from the Field, *Journal of Economic Literature* **47**(2), 315–372.

Dennis, K.: 1982a, Economic Theory and the Problem of Translation, *Journal of Economic Issues* **16**(3), 691–712.

Dennis, K.: 1982b, Economic Theory and the Problem of Translation: Part Two, *Journal of Economic Issues* **16**(4), 1039–1062.

Dennis, K.: 1996, A Logical Critique of Mathematical Formalism in Economics, *Journal of Economic Methodology* **3**(1), 151–169.

Dennis, K.: 1998a, Introduction, *in* K. Dennis (ed.), *Rationality in Economics: Alternative Perspectives*, Springer, New York, pp. 1–4.

Dennis, K.: 1998b, Rationality: A Journey Through the Semantic Bog, *in* K. Dennis (ed.), *Rationality in Economics: Alternative Perspectives*, Springer, New York, pp. 79–110.

Diamond, P. A.: 1967, Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility, *The Journal of Political Economy* **75**(5), 765–766.

Diecidue, E. and Wakker, P. P.: 2001, On the Intuition of Rank-Dependent Utility, *Journal of Risk and Uncertainty* **23**(3), 281–298.

Dietrich, F. and List, C.: 2013, A Reason-based Theory of Rational Choice, *Nous* **47**(1), 104–134.

Dietrich, F. and List, C.: 2016, Reason-based Choice and Context-dependence: An Explanatory Framework, *Economics and Philosophy* (Forhcoming).

Dilworth, C.: 1988, Identity, Equality and Equivalence, *Dialectica* **42**(2), 83–92.

Doyle, J. R.: 2013, Survey of Time Preference, Delay Discounting Models, *Judgment and Decision Making* **8**(2), 116–135.

Dreber, A., Ellingsen, T., Johannesson, M. and Rand, D. G.: 2013, Do People Care about Social Context? Framing Effects in Dictator Games, *Experimental Economics* **16**(3), 349–371.

Dreber, A., Fudenberg, D., Levine, D. K. and Rand, D. G.: 2016, Self-Control, Social Preferences and the Effect of Delayed Payments, *Working Paper* .
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477454*

Dufwenberg, M.: 2008, Psychological Games, *in* S. N. Durlauf and L. E. Blume (eds), *The New Palgrave Dictionary of Economics. Second Edition. Volume 6.*, Palgrave, Hampshire, pp. 714–717.

Dufwenberg, M., Gächter, S. and Hennig-Schmidt, H.: 2011, The Framing of Games and the Psychology of Play, *Games and Economic Behavior* **73**(2), 459–478.

Eckel, C. and Gintis, H.: 2010, Blaming the Messenger: Notes on the Current State of Experimental Economics, *Journal of Economic Behavior & Organization* **73**(1), 109–119.

Ege, R., Igersheim, H. and Le Chapelain, C.: 2012, Par-delà le transcendantal et le comparatif: deux arguments, *Revue française d'économie* **XXVII**(4), 185.

Egidi, M.: 2008, Le processus dual du raisonnement: origines, problèmes et perspectives, *in* B. Walliser (ed.), *Economie et Cognition*, Editions de la Maison des sciences de l'homme et Editions Ophrys, Paris, pp. 11–54.

Egidi, M.: 2012, The Cognitive Explanation of Economic Behaviour: From Simon to Kahneman, *in* R. Arena, A. Festré and N. Lazaric (eds), *The Handbook of Economics and Knowledge*, Edward Elgar Publishing, Cheltenham, pp. 183–210.

Ellingsen, T.: 1994, Cardinal Utility: A History of Hedonimetry, *in* M. Allais and O. Hagen (eds), *Cardinalism*, Springer, Dordrecht, pp. 105–166.

Elster, J.: 1979, *Ulysses and the Sirens: Studies in Rationaltiy and Irrationality*, Cambridge University Press, Cambridge, England.

Engel, C.: 2011, Dictator Games: A Meta Study, *Experimental Economics* **14**(4), 583–610.

Engelmann, D. and Hollard, G.: 2010, Reconsidering the Effect of Market Experience on the "Endowment Effect", *Econometrica* **78**(6), 2005–2019.

English, M. C. W. and Visser, T. A. W.: 2014, Exploring the repetition paradox: The effects of learning context and massed repetition on memory, *Psychonomic bulletin & review* **21**(4), 1026–1032.

Epper, T. and Fehr-Duda, H.: 2015, The Missing Link: Unifying Risk Taking and Time Discounting, *Working Paper* .
URL: *http://thomasepper.com/papers/wp/missing_link.pdf*

Evans, J. S. B.: 2004, History of Dual Process Theory of Reasoning, *in* K. Manktelow and M. Cheung Chung (eds), *Psychology of Reasoning: Theoretical and Historical Perspectives*, Psychology Press, Hove and New York, pp. 241–266.

Evans, J. S. B.: 2008, Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition, *Annual review of Psychology* **59**, 255–278.

Evans, J. S. B. and Frankish, K.: 2009, The Duality of Mind: An Historical Perspective, *in* J. S. B. Evans and K. Frankish (eds), *In Two Minds: Dual Processes and Beyond*, Oxford University Press, Oxford, pp. 1–30.

Evans, J. S. B. and Stanovich, K. E.: 2013a, Dual-Process Theories of Higher Cognition: Advancing the Debate, *Perspectives on Psychological Science* **8**(3), 223–241.

Evans, J. S. B. and Stanovich, K. E.: 2013b, Theory and Metatheory in the Study of Dual Processing: Reply to Comments, *Perspectives on Psychological Science* **8**(3), 257–262.

Fagley, N. S.: 1993, A Note Concerning Reflection Effects Versus Framing Effects, *Psychological Bulletin* **113**(3), 451–452.

Faralla, V., Benuzzi, F., Nichelli, P. and Dimitri, N.: 2012, Gains and Losses in Intertemporal Preferences: A Behavioral Study, *in* A. Innocenti and A. Sirigu (eds), *Neuroscience and the Economics of Decision Making*, Routledge, London and New York.

Farell, J. and Rabin, M.: 1996, Cheap Talk, *The Journal of Economic Perspectives* **10**(3), 103–118.

Favereau, J.: 2014, L'approche expérimentale du J-Pal en économie du développement : un tournant épistémologique?, *PhD Dissertation (Université Paris 1 Panthéon-Sorbonne)* .

Fehr-Duda, H. and Epper, T.: 2012, Probability and Risk: Foundations and Economic Implications of Probability- Dependent Risk Preferences, *Annual Review of Economics* **4**(1), 567–593.

Fehr, E. and Schmidt, K. M.: 1999, A Theory of Fairness, Competition, and Cooperation, *The Quarterly Journal of Economics* **114**(3), 817–868.

Fehr, E. and Schmidt, K. M.: 2010, On Inequity Aversion: A Reply to Binmore and Shaked, *Journal of Economic Behavior & Organization* **73**(1), 101–108.

Fennema, H. and Wakker, P. P.: 1997, Original and Cumulative Prospect Theory: A Discussion of Empirical Differences, *Journal of Behavioral Decision Making* **10**(1), 53–64.

Festré, A. and Garrouste, P.: 2008, Rationality, Behavior, Institutional, and Economic Change in Schumpeter, *Journal of Economic Methodology* **15**(4), 365–390.

Fishburn, P. C.: 1970, *Utility Theory for Decision Making*, John Wiley & Sons, Inc., New York.

Fishburn, P. C.: 1989a, Foundations of Decision Analysis: Along the Way, *Management Science* **35**(4), 387–406.

Fishburn, P. C.: 1989b, Retrospective on the Utility Theory of von Neumann and Morgenstern, *Journal of Risk and Uncertainty* **2**(2), 127–158.

Fishburn, P. C. and Rubinstein, A.: 1982, Time Preference, *International Review of Economics* **23**(3), 677–694.

Fishburn, P. C. and Wakker, P. P.: 1995, The Invention of the Independence Condition for Preferences, *Management Science* **41**(7), 1130–1144.

Fontaine, P.: 2000, Making Use of the Past: Theorists and Historians on the Economics of Altruism, *The European Journal of the History of Economic Thought* **7**(3), 407–422.

Frankish, K.: 2010, Dual-Processes and Dual System Theories of Reasoning, *Philosophy Compass* **5**(10), 914–926.

Frantz, R.: 2005, *Two Minds. Intuition and Analysis in the History of Economic Thought*, Springer, New York.

Frederick, S.: 2005, Cognitive Reflection and Decision Making, *The Journal of Economic Perspectives* **19**(4), 25–42.

Frederick, S., Loewenstein, G. and O'Donoghue, T.: 2002, Time Discounting and Time Preference: A Critical Review, *Journal of Economic Literature* **40**(2), 351–401.

Frey, B. S. and Stutzer, A.: 2007, Economics and Psychology: Developments and Issues, *in* B. S. Frey and A. Stutzer (eds), *Economics and Psychology: A Promising New Cross-Disciplinary Field*, The MIT Press, Cambridge, MA and London, pp. 3–15.

Friedman, M.: 1953, The Methodology of Positive Economics, *Essays in Positive Economics*, The University of Chicago Press, Chicago and London, pp. 3–45.

Frisch, D.: 1993, Reasons for Framing Effects, *Organizational Bheavior and Human Decision Processes* **54**, 399–429.

Fudenberg, D.: 2006, Advancing Beyond Advances in Behavioral Economics, *Journal of Economic Literature* **XLIV**(September), 694–711.

Fudenberg, D. and Levine, D. K.: 2006, A Dual-Self Model of Impulse Control, *The American Economic Review* **96**(5), 1449–1476.

Fudenberg, D. and Levine, D. K.: 2009, *A Long-run Collaboration on Long-run Games*, World Scientific Publishing, Singapore.

Fudenberg, D. and Levine, D. K.: 2011, Risk, Delay, and Convex Self-control Costs, *American Economic Journal: Microeconomics* **3**(3), 34–68.

Fudenberg, D. and Levine, D. K.: 2012a, Fairness, Risk Preferences and Independence: Impossibility Theorems, *Journal of Economic Behavior & Organization* **81**(2), 606–612.

Fudenberg, D. and Levine, D. K.: 2012b, Supplement to "Timing and Self-Control", *Econometrica* **80**(Online), 1–16.

Fudenberg, D. and Levine, D. K.: 2012c, Timing and Self-Control, *Econometrica* **80**(1), 1–42.

Fudenberg, D., Levine, D. K. and Maniadis, Z.: 2014, An Approximate Dual-self Model and Paradoxes of Choice Under Risk, *Journal of Economic Psychology* **41**(April), 55–67.

Gabbay, D. M. and Woods, J.: 2005, The Practical Turn in Logic, *in* D. M. Gabbay and F. Guenthner (eds), *Handbook of Philosophical Logic, Volume 13*, 2 edn, Springer, pp. 15–122.

Gajdos, T. and Weymark, J. A.: 2012, Introduction to Inequality and Risk, *Journal of Economic Theory* **147**(4), 1313–1330.

Galison, P.: 1999, Trading Zone: Coordinating Action and Belief, *in* M. Biagioli (ed.), *The Science Studies Reader*, Routledge, New York and London, pp. 137–160.

Geiger, N.: 2016, The Rise of Behavioural Economics: A Quantitative Asessment, *Working Paper* .
URL: *https://ideas.repec.org/p/zbw/hohpro/442015.html*

Geurts, B.: 2013, Alternatives in Framing and Decision Making, *Mind & Language* **28**(1), 1–19.

Gilardone, M.: 2010, Amartya Sen sans prisme, *Cahiers d'économie Politique/Papers in Political Economy* **58**(1), 9–39.

Gilboa, I., Maccheroni, F., Marinacci, M. and Schmeidler, D.: 2010, Objective and Subjective Rationality in a Multiple Prior Model, *Econometrica* **78**(2), 755–770.

Giocoli, N.: 2003, *Modeling Rational Agents: From Interwar Economics to Early Modern Game Theory*, Edward Elgar Publishing, Cheltenham.

Giraud, G. and Renouard, C.: 2011, Is the Veil of Ignorance Transparent?, *Oeconomia* **1**(2), 239–258.

Giraud, R.: 2004a, Framing under risk: Endogenizing the Reference Point and Separating Cognition and Decision, *Working Paper* .
URL: *http://econpapers.repec.org/paper/msewpsorb/bla04090.htm*

Giraud, R.: 2004b, *Une théorie de la décision pour les préférences imparfaites*, PhD thesis.

Giraud, R.: 2005, Anomalies de la théorie des préférence. Une interprétation et une proposition de formalisation, *Revue Economique* **56**(4), 829–854.

Giraud, R.: 2012, Money Matters: An Axiomatic Theory of the Endowment Effect, *Economic Theory* **50**(2), 303–339.

Glimcher, P. W. and Fehr, E.: 2014, Introduction: A Brief History of Neuroeconomics, *in* P. W. Glimcher and E. Fehr (eds), *Handbook of Neuroeconomics (2nd ed.)*, Elsevier, New York, pp. xvii–xxviii.

Glimcher, P. W., Kable, J. and Louie, K.: 2007, Neuroeconomic Studies of Impulsivity: Now or Just as Soon as Possible?, *The American Economic Review* **97**(2), 142–147.

Gold, N. and List, C.: 2004, Framing as Path Dependence, *Economics and Philosophy* **20**(2), 253–277.

Goldman, S. M.: 1980, Consistent Plans, *The Review of Economic Studies* **47**(3), 533–537.

Gollier, C.: 2001, *The Economics of Risk and Time*, The MIT Press, Cambridge, MA and London.

Gorman, W. M.: 1968, The Structure of Utility Functions, *The Review of Economic Studies* **35**(4), 367–390.

Grant, S. and van Zandt, T.: 2009, Expected Utility Theory, *in* P. Anand, P. K. Pattanaik and C. Puppe (eds), *The Handbook of Rational and Social Choice: An Overview of New Foundations and Applications*, Oxford University Press, Oxford, pp. 21–68.

Green, M.: 2014, Speech Acts, *Stanford Encyclopedia of Philosophy* .
URL: *http://plato.stanford.edu/entries/speech-acts/*

Grice, P. H.: 1957, Meaning, *The Philosophical Review* **66**(3), 377–388.

Grice, P. H.: 1975, Logic and Conversation, *in* P. Cole and J. L. Morgan (eds), *Syntax and Semantics 3: Speech Acts*, Elsevier, New York, pp. 41–58.

Gu, Y.: 1993, The Impasse of Perlocution, *Journal of Pragmatics* **20**(5), 405–432.

Guala, F.: 2000, The Logic of Normative Falsification: Rationality and Experiments in Decision Theory, *Journal of Economic Methodology* **7**(1), 59–93.

Guala, F.: 2005, *The Methodology of Experimental Economics*, Cambridge University Press, Cambridge, England.

Guala, F.: 2012, Are Preferences for Real? Choice Theory, Folk Psychology, and the Hard Case for Commonsensible Realism, *in* A. Lehtinen, J. Kurokikoski and P. Ylikoski (eds), *Economics for Real: Uskali Mäki and the Place of Truth in Economics*, Routledge, Abingdon, pp. 137–154.

Gul, F.: 1992, Savage's Theorem with a Finite Number of States, *Journal of Economic Theory* **57**(1), 99–110.

Gul, F. and Pesendorfer, W.: 2001, Temptation and Self-control, *Econometrica* **69**(6), 1403–1435.

Gul, F. and Pesendorfer, W.: 2004a, Self-control and the Theory of Consumption, *Econometrica* **7**(1), 119–158.

Gul, F. and Pesendorfer, W.: 2004b, Self-control, Revealed Preference and Consumption Choice, *Review of Economic Dynamics* **7**(2), 243–264.

Gul, F. and Pesendorfer, W.: 2005, The Revealed Preference Theory of Changing Tastes, *Review of Economic Studies* **72**(2), 429–448.
URL: *http://restud.oxfordjournals.org/lookup/doi/10.1111/j.1467-937X.2005.00338.x*

Gul, F. and Pesendorfer, W.: 2007, Harmful Addiction, *The Review of Economic Studies* **74**(1), 147–172.

Gul, F. and Pesendorfer, W.: 2008, The Case for Mindless Economics, *in* A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics*, Oxford University Press, New York, pp. 3–40.

Güth, W., Levati, V. and Ploner, M.: 2008, On the Social Dimension of Time and Risk Preferences: An Experimental Study, *Economic Inquiry* **46**(2), 261–272.

Halevy, Y.: 2008, Strotz Meets Allais: Diminishing Impatience and the Certainty Effect, *American Economic Review* **98**(3), 1145–1162.

Hammond, P. J.: 1976, Changing Tastes and Coherent Dynamic Choice, *Review of Economic Studies* **43**(1), 159–173.

Hammond, P. J.: 1977, Dynamic Restrictions on Metastatic Choice, *Economica* **44**(176), 337–350.

Hammond, P. J.: 1983, Ex-post Optimality as a Dynamically Consistent Objective for Collective Choice Under Uncertainty, *in* P. K. Pattanaik and M. Salles (eds), *Social Choice and Welfare*, North-Holland, Amsterdam, New York and Oxford, pp. 175–205.

Hammond, P. J.: 1987, Altruism, *in* J. M. Eatwell and P. Newman (eds), *The New Palgrave Dictionary of Economics*, Palgrave, London, pp. 85–87.

Hammond, P. J.: 1988a, Consequentialism and the Independence Axiom, *in* B. Munier (ed.), *Risk, Decision and Rationality*, Reidel, Dordrecht, pp. 503–516.

Hammond, P. J.: 1988b, Consequentialist Foundations for Expected Utility Theory, *Theory and Decision* **25**, 25–78.

Hammond, P. J.: 1989, Consistent Plans, Consequentialism, and Expected Utility, *Econometrica* **57**(6), 1445–1449.

Hammond, P. J.: 1998, Objective expected utility, *in* S. Barberà, P. J. Hammond and C. Seidl (eds), *Handbook of utility Theory*, Kluwer, Dordrecht, chapter 5.

Hammond, P. J. and Zank, H.: 2014, Rationality and Dynamic Consistency Under Risk and Uncertainty, *in* M. J. Machina and W. K. Viscusi (eds), *Handbook of the Economics of Risk and Uncertainty*, Elsevier, Amsterdam, pp. 41–98.

Hands, W. D.: 2001, *Reflections Wihout Rules*, Cambridge University Press, Cambridge.

Hands, W. D.: 2012a, Realism, Commonsensibles, and Economics: The Case of Contemporary Revealed Prefenrence Theory, *in* A. Lehtinen, J. Kuorikoski and P. Ylikoski (eds), *Economics for Real: Uskali Mäki and the Place of Truth in Economics*, Routledge, Abingdon, pp. 1–29.

Hands, W. D.: 2012b, The Positive-Normative Dichotomy and Economics, *in* U. Mäki (ed.), *Handbook of the Philosophy of Science, Vol.5: Philosophy of Economics*, Elsevier, Oxford, pp. 219–239.

Hands, W. D.: 2013a, Foundations of Contemporary Revealed Preference Theory, *Erkenntnis* **78**(5), 1081–1108.

Hands, W. D.: 2013b, Normative Rational Choice Theory: Past, Present, and Future, *Working Paper* .
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1738671*

Hands, W. D. and Mirowski, P. E.: 1998, Harold Hotelling and The Neoclassical Dream, *in* R. E. Backhouse, D. M. Hausman, U. Maki and A. Salanti (eds), *Economics and Methodology: Crossing Boundaries*, St. Martin's Press, New York, pp. 322–397.

Hanemann, W. M.: 1991, Willingness to Pay and Willingness to Accept: How Much Can They Differ?, *American Economic Review* **81**(3), 635–47.

Harris, C. and Laibson, D.: 2013, Instantaneous Gratification, *The Quarterly Journal of Economics* **128**(1), 205–248.

Harrison, G. W., Lau, M. I. and Rutström, E. E.: 2007, Estimating Risk Attitudes in Denmark: A Field Experiment, *The Scandinavian Journal of Economics* **109**(2), 341–368.

Harrison, G. W. and Rutström, E. E.: 2008, Risk Aversion in the Laboratory, *in* J. C. Cox and G. W. Harrison (eds), *esearch in Experimental Economics, Volume 12*, Emerald Group Publishing, Bingley, pp. 41–196.

Harrison, G. W. and Rutström, E. E.: 2009, Expected Utility Theory and Prospect Theory: One Wedding and a Decent Funeral, *Experimental Economics* **12**(2), 133–158.

Harsanyi, J. C.: 1955, Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility, *The Journal of Political Economy* **63**(4), 309–321.

Harsanyi, J. C.: 1988, Assessing Other People's Utilities, *in* B. R. Munier (ed.), *Risk, Decision and Rationality*, Reidel, Dordrecht, pp. 127–138.

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*, Oxford University Press, New York, pp. 125–153.

Hausman, D. M.: 2012, *Preference, Value, Choice, and Welfare*, Cambridge University Press, New York.

Hausman, D. M. and Mcpherson, M. S.: 2006, *Economic Analysis, Moral Philosophy, and Public Policy (Second Edition)*, Cambridge University Press, Cambridge, England.

Hédoin, C.: 2016, Sen's Critique of Revealed Preferences Theory and Its "Neo-Samulsonian" Critique: A Methodological and Theoretical Asessment, *Journal of Economic Methodology* **Forthcomin**.

Herfeld, C. S.: 2013, Axiomatic Choice Theory Traveling between Mathematical Formalism, Normative Choice Rules and Psychological Measurement, 1944-1956, *Center for the History of Political Economy Working Paper* .
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2296884*

Herfeld, C. S.: 2016, Defining the Rules of Rationality: Marschak, Koopmans, and the Normative Shift in Economics, 1943-1954, *History of Political Economy* (Forthcoming).

Heukelom, F.: 2009, *Kahneman and Tversky and the Making of Behavioral Economics (phD thesis)*, PhD thesis, Amsterdam.

Heukelom, F.: 2012, A Sense of Mission: The Alfred P. Sloan and Russell Sage Foundations' Behavioral Economics Program, 1984-1992, *Science in Context* **2**(25), 263–286.

Heukelom, F.: 2014, *Behavioral Economics: A history*, Oxford University Press, New York.

Higgins, T. E.: 2012, Motivation Science in Social Psychology: A Tale of Two Histories, *in* A. W. Kruglanski and W. Stroebe (eds), *Handbook of the History of Social Psychology*, Psychology Press, New York and London, pp. 199–218.

Hilton, D. J.: 1995, The Social Context of Reasoning: Conversational Inference and Rational Judgment, *Psychological Bulletin* **118**(2), 248–271.

Hilton, D. J.: 2011, Linguistic Polarity, Outcome Framing, and the Structure of Decision Making: A Pragmatic Approach, *in* G. Keren (ed.), *Perspectives on Framing*, Psychology Press, New York, pp. 135–156.

Hoffman, E., McCabe, K. and Smith, V. L.: 1996, Social Distance and Other-regarding Behavior in Dictator Games, *The American Economic Review* **86**(3), 653–660.

Hollard, G.: 2016, Consistent Inconsistencies? Evidence from Decision Under Risk, *Theory and Decision* **80**(4), 623–648.

Holt, C. A. and Laury, S. K.: 2014, Assessment and Estimation of Risk Preferences, *in* M. J. Machina and W. K. Viscusi (eds), *Handbook of the Economics of Risk and Uncertainty*, Elsevier, Amsterdam, pp. 135–202.

Horn, L. R.: 2006, Implicature, *in* L. R. Horn and G. Ward (eds), *The Handbook of Pragmatics*, Blackwell, Malden, pp. 3–28.

Huutoniemi, K., Klein, J. T., Bruun, H. and Huukkinen, J.: 2010, Analyzing Interdisciplinarity: Typology and Indicators, *Research Policy* **39**(1), 79–88.

Ida, T. and Ogawa, K.: 2013, Inequality Aversion, Social Discount, and Time Discount Rates, *Working Paper* .
URL: *https://www.econ.kyoto-u.ac.jp/projectcenter/Paper/e-10-013.pdf*

Igou, E.: 2011, The When and Why of Risky Choice Framing Effects: A Constructive Processing Perspective, *in* G. Keren (ed.), *Perspectives on Framing*, Psychology Press, New York, pp. 219–238.

Isoni, A., Loomes, G., Sugden, R., Plott, C. R. and Zeiler, K.: 2011, The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: Comment, *American Economic Review* **101**(2), 1012–1028.

Israel, M.: 2006, The Pragmatics of Polarity, *in* L. R. Horn and G. Ward (eds), *The Handbook of Pragmatics*, Blackwell, Malden, pp. 701–723.

Jallais, S. and Pradier, P.-C.: 2005, The Allais Paradox and its Immediate Consequences for Expected Utility Theory, *in* P. Fontaine and R. J. Leonard (eds), *The Experiment in the History of Economics*, Routledge, London and New York, pp. 21–42.

Jallais, S., Pradier, P.-C. and David, T.: 2008, Facts, Norms and Expected Utility Functions, *History of the Human Sciences* **21**(2), 45–62.

Jamison, J., Karlan, D. and Zinman, J.: 2012, Measuring Risk and Time Preferences and Their Connections with Behavior, *Working Paper* .
URL: *http://www.dartmouth.edu/ jzinman/Papers/Measuring Risk and Time Prefs.pdf*

Jolls, C., Sunstein, C. R. and Thaler, R. H.: 1998, A Behavioral Approach to Law and Economics, *Stanford Law review* (50), 1471–1550.

Jones, B. and Rachlin, H.: 2006, Social Discounting, *Psychological Science* **17**(4), 283–286.

Judd, C. M. and Brauer, M.: 1995, Repetition and Evaluative Extremity, *in* R. E. Petty and J. A. Krosnick (eds), *Attitude Strength: Antecedents and Consequences*, Psychology Press, New Haven & London, pp. 43–72.

Juille, T.: 2016, Consumer Sovereignty and Consumer Identity, *Working Paper* .

Jullien, D.: 2013, Intentional Apple-Choice Behaviors: When Amartya Sen Meets John Searle, *Cahiers d'économie Politique/Papers in Political Economy* **65**, 97–128.

Jullien, D.: 2016a, All Frames Created Equal are not Identical: On the Structure of Kahneman and Tversky's Framing Effects, *Oeconomia* (Forthcoming).

Jullien, D.: 2016b, Corrigendum et commentaires sur "Le paradoxe d'Allais. Comment lui rendre sa signification perdue?" de Philippe Mongin [2014], *Revue Economique* (Forthcoming).

Jullien, D. and Vallois, N.: 2014, A Probabilistic Ghost in the Experimental Machine, *Journal of Economic Methodology* **21**(3), 232–250.

Jurdant, B.: 1973, *Les problèmes théoriques de la vulgarisation scientifique*, Ph.d, Université Louis Pasteur de Strasbourg.

Kahneman, D.: 1994, New Challenges to the Rationality Assumption, *Journal of Institutional and Theoretical Economics* **150**(1), 18–36.

Kahneman, D.: 2000, Preface, *in* A. Tversky and D. Kahneman (eds), *Choices, Values and Frames*, Cambridge University Press, New York.

Kahneman, D.: 2003, A Psychological Perspective on Economics, *The American Economic Review* **93**(2), 162–168.

Kahneman, D.: 2011, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.

Kahneman, D., Knetsch, J. L. and Thaler, R.: 1986a, Fairness as a Constraint on Profit Seeking: Entitlements in the Market, *American Economic Review* **76**(4), 728–741.

Kahneman, D., Knetsch, J. L. and Thaler, R. H.: 1986b, Fairness and the Assumptions of Economics, *The Journal of Business* **59**(4), S285–S300.

Kahneman, D., Knetsch, J. L. and Thaler, R. H.: 1990, Experimental Tests of the Endowment Effect and the Coase Theorem, *Journal of Political Economy* **98**(6), 1325.

Kahneman, D., Knetsch, J. L. and Thaler, R. H.: 1991, Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias, *Journal of Economic Perspectives* **5**(1), 193–206.

Kahneman, D. and Sugden, R.: 2005, Experienced Utility as a Standard of Policy Evaluation, *Environmental and Resource Economics* **32**(1), 161–181.

Kahneman, D. and Thaler, R. H.: 2006, Anomalies: Utility Maximization and Experienced Utility, *Journal of Economic Perspectives* **20**(1), 221–234.

Kahneman, D. and Tversky, A.: 1979, Prospect Theory: An Analysis of Decision under Risk, *Econometrica* **47**(2), 263–291.

Kahneman, D. and Tversky, A.: 1984, Choices, Values, and Frames, *American Psychologist* **39**(4), 341–350.

Kahneman, D., Wakker, P. P. and Sarin, R.: 1997, Back to Bentham? Explorations of Experienced Utility, *Quarterly Journal of Economics* **112**(2), 375–405.

Kang, Q.: 2013, On Perlocutionary Act, *Studies in Literature and Language* **6**(1), 60–64.

Kao, Y.-F. and Velupillai, V.: 2015, Behavioural economics: classical and modern, *The European Journal of the History of Economic Thought* **22**(223-271).

Keren, G.: 2011a, On the Definition and Possible Underpinnings of Framing Effects: A Brief Review and a Critical Evaluation, *in* G. Keren (ed.), *Perpectives on Framing*, Psychology Press, New York, chapter 1, pp. 3–34.

Keren, G.: 2011b, *Perpectives on Framing*, Psychology Press, New York.

Keren, G.: 2012, Framing and Communication: The Role of Frames in Theory and in Practice, *Netspar Panel Papers* **32**(Octobre), 11–58.

Keren, G.: 2013, A Tale of Two Systems: A Scientific Advance or A Theoretical Stone Soup?, *Perspectives on Psychological Science* **8**(3), 257–262.

Keren, G. and Roelofsma, P.: 1995, Immediacy and Certainty in Intertemporal Choice, *Organizational Behavior and Human Decision Processes* **63**(3), 287–297.

Keren, G. and Schul, Y.: 2009, Two Is Not Always Better Than One: A Critical Evaluation of Two-System Theories, *Perspectives on Psychological Science* **4**(6), 533–550.

Keysar, B., Hayakawa, S. and An, S. G.: 2012, The Foreign-language Effect: Thinking in a Foreign Tongue Reduces Decision Biases, *Psychologican Science* **23**(6), 661–668.

Kirby, K. N. and Guastello, B.: 2001, Making Choices in Anticipation of Similar Future Choices can Increase Self-control, *Journal of Experimental Psychology: Applied* **7**(2), 154–164.

Kirman, A.: 1989, The Intrinsic Limits of Modern Economic Theory: The Emperor Has no Clothes, *The Economic Journal* **99**(395), 126–139.

Kitch, E. W.: 1990, The Framing Hypothesis: Is It Supported by Credit Card Issuer Opposition to a Surcharge on a Cash Price?, *Journal of Law, Economics and Organization* **6**(1), 217–33.

Kitcher, P.: 1984, 1953 and all that. A Tale of Two Sciences, *The Philosophical Review* **93**(3), 335–373.

Klein, J. T.: 2010, A Taxonomy of Interdiscinarity, *in* R. Frodeman (ed.), *The Oxford Handbook of Interdisciplinarity*, Oxford University Press, Oxford, pp. 15–31.

Koopmans, T. C.: 1960, Stationary Ordinal Utility and Impatience, *Econometrica* **28**(2), 287–309.

Korta, K. and Perry, J.: 2015, Pragmatics, *Stanford Encyclopedia of Philosophy* . **URL:** *http://plato.stanford.edu/entries/pragmatics/*

Köszegi, B.: 2010, Utility from Anticipation and Personal Equilibrium, *Economic Theory* **44**(3), 415–444.

Köszegi, B. and Rabin, M.: 2006, A Model of Reference-Dependent Preferences, *The Quarterly Journal of Economics* **121**(4), 1133–1165.

Koszegi, B. and Rabin, M.: 2007, Reference-Dependent Risk Attitudes, *The American Economic Review* **97**(4), 1047–1073.

Köszegi, B. and Rabin, M.: 2008a, Choices, Situations, and Happiness, *Journal of Public Economics* **92**(8-9), 1821–1832.

Köszegi, B. and Rabin, M.: 2008b, Revealed Mistakes and Revealed Preferences, *in* A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics: A Handbook*, Oxford University Press, New York, pp. 193–209.

Köszegi, B. and Rabin, M.: 2009, Reference-Dependent Consumption Plans, *American Economic Review* **99**(3), 909–936.

Kovarik, J.: 2009, Giving it Now or Later: Altruism and Discounting, *Economics Letters* **102**(3), 152–154.

Krantz, D. H., Luce, D. R., Suppes, P. and Tversky, A.: 1971, *Foundations of Measurement Volume I: Additive and Polynomial Representations*, Academic Press, San Diego and London.

Krawczyk, M. and Le Lec, F.: 2010, 'Give Me a Chance!' An Experiment in Social Decision Under Risk, *Experimental Economics* **13**(4), 500–511.

Kreiner, H. and Gamliel, E.: 2016, Looking at Both Sides of the Coin: Mixed Representation Moderates Attribute framing Bias in Written and Auditory Messages, *Applied Cognitive Psychology* (Advanced Online Publication).

Kreps, D.: 1988, *Notes on the Theory of Choice*, Westview Press, Boulder.

Kreps, D. and Porteus, E.: 1978, Temporal Resolution of Uncertainty and Dynamic Choice Theory, *Econometrica* **46**(1), 185–200.

Kruglanski, A. W.: 2013, Only One? The Default Interventionist Perspective as a Unimodel, *Perspectives on Psychological Science* **8**(3), 242–247.

Kühberger, A.: 1995, The Framing of Decisions: A New Look at Old Problems, *Organizational Bheavior and Human Decision Processes* **62**(2), 230–240.

Kühberger, A.: 1997, Theoretical conceptions of framing effects in risky decisions, *in* R. Ranyard, R. W. Crozier and O. Svenson (eds), *Decision Making. Cognitive models and explanations*, Routledge, London and New-york, chapter 8, pp. 128–144.

Kühberger, A.: 1998, The Influence of Framing on Risky Decisions: A Meta-analysis, *Organizational Behavior and Human Decision Processes* **75**(1), 23–55.

Kühberger, A. and Gradle, P.: 2013, Choice, Rating, and Ranking: Framing Effects with Different Response Modes, *Journal of Behavioral Decision Making* **26**(2), 109–117.

Kühberger, A., Schulte-Mecklenbeck, M. and Perner, J.: 1999, The Effects of Framing, Reflection, Probability, and Payoff on Risk Preference in Choice Tasks., *Organizational behavior and human decision processes* **78**(3), 204–231.

Kühberger, A., Schulte-Mecklenbeck, M. and Perner, J.: 2002, Framing Decisions: Hypothetical and Real, *Organizational Behavior and Human Decision Processes* **89**(2), 1162–1175.

Kühberger, A. and Tanner, C.: 2010, Risky Choice Framing: Task Versions and a Comparison of Prospect Theory and Fuzzy-Theory, *Journal of Behavioral Decision Making* **23**(3), 314–329.

Kühberger, A. and Wiener, C.: 2012, Explaining Risk Attitude in Framing Tasks by Regulatory Focus: A Verbal Protocol Analysis and a Simulation Using Fuzzy Logic, *Decision Analysis* **9**(4), 359–372.

Kuhl, B. A. and Anderson, M. C.: 2011, More is not always better: paradoxical effects of repetition on semantic accessibility, *Psychonomic bulletin & review* **18**(5), 964–972.

Kurzon, D.: 1998, The Speech Act Status of Incitement: Perlocutionary Acts Revisited, *Journal of Pragmatics* **29**(5), 571–596.

Laibson, D.: 1994, Hyperbolic Discounting and Consumption, (Ph.D dissertation, MIT).

Laibson, D.: 1997, Golden Eggs and Hyperbolic Discounting, *The Quarterly Journal of Economics* **112**(2), 443–478.

Lallement, J.: 1984, *Les Fondements de la Théorie Néoclassique de la Demande*, PhD thesis, Paris 1 Panthéon-Sorbonne.

Lanzi, D.: 2011, Frames as Choice Superstructures, *The Journal of Socio-Economics* **40**(2), 115–123.

Lapied, A. and Renault, O.: 2012, An Investigation of Time Consistency for Subjective Discontinued Utility, *Working Paper* .
URL: *https://halshs.archives-ouvertes.fr/halshs-00793174/*

LeBoeuf, R. A. and Shafir, E.: 2003, Deep Thoughts and Shallow Frames: On the Susceptibility to Framing Effects, *Journal of Behavioral Decision Making* **16**(2), 77–92.

Lecouteux, G.: 2015, Reconciling Normative and Behavioural Economics, *PhD Dissertation* (Ecole Polytechnique).

Lee, K. S.: 2013, The Legitimization of Laboratory Experiments in the Economics Profession during the Reagan Era, *Working Paper* .
URL: *http://econ.duke.edu/uploads/media_items/lee-2013-hope-lunch.original.pdf*

Leliveld, M. C., van Beest, I., van Dijk, E. and Tenbrunsel, A. E.: 2009, Understanding the Influence of Outcome Valence in Bargaining: A Study on Fairness Accessibility, Norms, and Behavior, *Journal of Experimental Social Psychology* **45**(3), 505–514.

Lerner, S.: 2014, A non-monotonic intensional framework for framing effects, *Journal of Economic Methodology* **21**(1), 37–53.

Levin, I. P.: 1987, Associative Effects of Information Framing, *Bulletin of the Psychonomic Society* **25**(2), 85–86.

Levin, I. P., Schneider, S. L. and Gaeth, G. J.: 1998, All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects., *Organizational behavior and human decision processes* **76**(2), 149–188.

Levine, D. K.: 2012, *Is Behavioral Economics Doomed?: The Ordinary versus the Extraordinary*, Open Book Publishers, Cambridge, England.
URL: *http://www.openbookpublishers.com/product/77*

Levitt, S. D. and List, J. A.: 2007a, Viewpoint: On the Generalizability of Lab Behaviour to the Field, *Canadian Journal of Economics/Revue canadienne d'économique* **40**(2), 347–370.

Levitt, S. D. and List, J. A.: 2007b, What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World ?, *Journal of Economic Perspectives* **21**(2), 153–174.

Levitt, S. D. and List, J. A.: 2008, Homo economicus evolves, *Science* **319**(5865), 909–10.

Liberman, V., Samuels, S. M. and Ross, L.: 2004, The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves, *Personality and Social Psychology Bulletin* **9**(9), 1175–1185.

Lichtenstein, S. and Slovic, P.: 2006, The Construction of Preference: An Overview, *in* S. Lichtenstein and P. Slovic (eds), *The Construction of Preference*, Cambridge University Press, New York, pp. 1–39.

Lipman, B. L.: 2003, Language and Economics, *in* N. Dimitri, M. Basili and I. Gilboa (eds), *Cognitive Processes and Economic Behaviour*, Routledge, London and New York.

Lipman, B. L. and Pesendorfer, W.: 2013, Temptation, *in* D. Acemoglu, M. Arellano and E. Dekel (eds), *Advances in Economics and Econometrics: Tenth World Congress, Volume 1*, Cambridge University Press, New York, pp. 243–288.

List, J. A.: 2007, On the Interpretation of Giving in Dictator Games, *Journal of Political Economy* **115**(3), 482–493.

Liu, F.: 2013, The Nature of Perlocution, *Journal of Cambridge Studies* **3**(1), 25–30.

Loewenstein, G. F.: 1988, Frames of Mind in Intertemporal Choice, *Management Science* **34**(2), 200–214.

Loewenstein, G. and Haisley, E.: 2008, The Economist as Therapist: Methodological Ramifications of "Light" Paternalism, *in* A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics: A Handbook*, Oxford University Press, New York, pp. 210–247.

Loewenstein, G., John, L. and Volpp, K. G.: 2013, Using Decision Errors to Help People Help Themselves, *in* E. Shafir (ed.), *The Behavioral Foundations of Public Policy*, Princeton University Press, Princeton and Oxford, pp. 361–379.

Loewenstein, G. and O'Donoghue, T.: 2004, Animal Spirits: Affective and Deliberative Processes in Economic Behavior, *Working Paper* .
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=539843*

Loewenstein, G. and O'Donoghue, T.: 2007, The Heat of the Moment : Modeling Interactions Between Affect and Deliberation, *Working Paper* .
URL: *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.335.5703&rep=rep1&type=pdf*

Loewenstein, G., O'Donoghue, T. and Bhatia, S.: 2015, Modeling the Interplay Between Affect and Deliberation, *Decision* **2**(2), 55–81.

Loewenstein, G., O'Donoghue, T. and Rabin, M.: 2003, Projection Bias in Predicting Future Utility, *The Quarterly Journal of Economics* **118**(4), 1209–1248.

Loewenstein, G. and Prelec, D.: 1992, Anomalies in Intertemporal Choice: Evidence and an Interpretation, *The Quarterly Journal of Economics* **107**(2), 573–597.

Loewenstein, G., Rick, S. and Cohen, J. D.: 2008, Neuroeconomics, *Annual Review of Psychology* **59**, 647–72.

Loewenstein, G. and Ubel, P. A.: 2008, Hedonic Adaptation and the Role of Decision and Experience Utility in Public Policy, *Journal of Public Economics* **92**(8-9), 1795–1810.

Luce, D. R.: 2000, *Utility of Gains and Losses: Measurement-theoretical and experimental approaches*, Lawrence Erlbaum Associates, Mahwah.

Luce, R. D. and Fishburn, P. C.: 1991, Rank- and Sign-dependent Linear Utility Models for Finite First-order Gambles, *Journal of Risk and Uncertainty* **4**(1), 29–59.

Luce, R. D., Suppes, P., Krantz, D. H. and Tversky, A.: 1990, *Foundations of Measurement. Volume III: Representation, Axiomatization, and Invariance*, Academic Press, New York.

MacCrimmon, K. R.: 1968, Descriptive and Normative Implications of the Decision Theory Postulates, *in* K. Borch and J. Mossin (eds), *Risk and Uncertainty*, MacMillan, Houndmills, pp. 3–23.

MacCrimmon, K. R. and Larson, S.: 1979, Utility Theory: Axioms vs Paradoxes, *in* M. Allais and O. Hagen (eds), *Expected Utility Hypothesis and the Allais Paradox*, Reidel, Dordrecht.

Machina, M. J.: 1987, Choice Under Uncertainty: Problems Solved and Unsolved, *Journal of Economic Perspectives* **1**(1), 121–154.

Machina, M. J.: 1989, Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty, *Journal of Economic Literature* **XXVII**(December), 1622–1668.

Machlup, F.: 1978, If Matter Could Talk, *Methodology of Economics and Other Social Sciences*, Academic Press, pp. 309–332.

Maclean, D.: 1985, Rationality and Equivalent Redescriptions, *in* M. Grauer, M. Thompson and A. Wierzbicki (eds), *Plural Rationality and Interactive Decision Processes*, Springer, Berlin, pp. 83–95.

Magen, E., Dweck, C. S. and Gross, J. J.: 2008, The Hidden-zero Effect: Representing a Single Choice as an Extended Sequence Reduces Impulsive Choice, *Psychological Science* **19**(7), 648–649.

Mäki, U.: 2013, Performativity: Saving Austin from Mackenzie, *in* V. Karakostas and D. Dieks (eds), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, Springer, Dordrecht, pp. 443–453.

Malinvaud, E.: 1952, Note on von Neumann-Morgenstern's Strong Independence Axiom, *Econometrica* **20**(4), 679.

Mandel, D. R.: 2014, Do Framing Effects Reveal Irrational Choice?, *Journal of Experimental Psychology: General* **143**(3), 1185–1198.

Mandel, D. R.: 2015, Communicating Numeric Quantities in Context: Implications for Decision Science and Rationality Claims, *Frontiers in psychology* **6**, 537–537.

Mandel, D. R. and Vartanian, O.: 2012, Frames, brains, and content domains: Neural and behavioral effects of descriptive context on preferential choice, *in* O. Vartanian and D. R. Mandel (eds), *Neuroscience of Decision Making*, Routledge, New York, pp. 45–68.

Manzini, P. and Mariotti, M.: 2014, A Case of Framing Effects: The Elicitation of Time Preferences, *Working Paper* .
URL: *http://repo.sire.ac.uk/handle/10943/587*

Marcu, D.: 2000, Perlocutions: The Achilles' Heel of Speech Act Theory, *Journal of Pragmatics* **32**(12), 1719–1741.

Markowitz, H.: 1952, The Utility of Wealth, *The Journal of Political Economy* **IX**(2), 151–158.

Martins, N. O.: 2013, The Place of the Capability Approach within Sustainability Economics, *Ecological Economics* **95**, 226–230.

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

Maule, J. and Villejoubert, G.: 2007, What lies beneath: Reframing framing effects, *Thinking & Reasoning* **13**(1), 25–44.

McCloskey, D. N.: 1998, *The Rhetoric of Economics. Second edition*, The University of Wisconsin Press, Madison.

McClure, S. M., Ericson, K. M., Laibson, D., Loewenstein, G. and Cohen, J. D.: 2007, Time Discounting for Primary Rewards, *The Journal of Neuroscience* **27**(21), 5796–804.

McClure, S. M., Laibson, D., Loewenstein, G. and Cohen, J. D.: 2004, Separate Neural Systems Value Immediate and Delayed Monetary Rewards, *Science* **306**(5695), 503–7.

McKenzie, C. R., Liersch, M. J. and Finkelstein, S. R.: 2006, Recommendations Implicit in Policy Defaults, *Psychological Science* **17**(5), 414–420.

McKenzie, C. R. M. and Nelson, J. D.: 2003, What a Speaker's Choice of Frame Reveals: Reference Points, Frame Selection, and Framing Effects, *Psychonomic bulletin & review* **10**(3), 596–602.

McNeil, B. J., Pauker, S. G., Harold C. Sox, J. and Tversky, A.: 1982, On the Elicitation of Preferences for Alternative Therapies, *New England Journal of Medicine* **306**, 1259–1262.

Meier, S.: 2007, A Survey of Economic Theories and Field Evidence on Pro-Social Behavior, *in* B. S. Frey and A. Stutzer (eds), *Economics and Psychology: A Promising New Cross-Disciplinary Field*, The MIT Press, Cambridge, MA and London, pp. 51–88.

Mellers, B. and Locke, C.: 2007, What have we learned from our mistakes?, *in* W. Edwards, R. F. J. Miles and D. von Winterfeldt (eds), *Advances in decision analysis. From foundations to applications*, Cambridge University Press, New York, pp. 351–374.

Meyer, J.: 2014, The Theory of Risk and Risk Aversion, *in* M. J. Machina and W. K. Viscusi (eds), *Handbook of the Economics of Risk and Uncertainty*, Elsevier, Amsterdam, pp. 99–134.

Meyerowitz, B. E. and Chaiken, S.: 1987, The Effect of Message Framing on Breast Self-examination: Attitudes, Intentions, and Behavior, *Journal of Personality and Social Psychology* **52**(3), 500–510.

Milanesi, J.: 2010, Éthique et évaluation monétaire de l'environnement : la nature est-elle soluble dans l'utilité ?, *VertigO* **10**(2).
URL: *http://vertigo.revues.org/10050*

Miller, P. M. and Fagley, N. S.: 1991, The Effects of Framing, Problem Variations, and Providing Rationale on Choice, *Personality and Social Psychology Bulletin* **17**(5), 517–522.

Mirowski, P.: 2002, *Machine Dreams: Economics Becomes a Cyborg Science*, Cambridge University Press, Cambridge, England.

Mirowski, P.: 2006, Twelve Theses Concerning the History of Postwar Neoclassical Price Theory, *History of Political Economy* **38**(annual suppl.), 343–379.

Mirowski, P.: 2009, Why There is (as yet) no Such Thing as an Economics of Knowledge, *in* H. Kincaid and D. Ross (eds), *The oxford Handbook of Philosophy of Economics*, Oxford University Press, New York, pp. 99–157.

Mirowski, P.: 2012, The Unreasonable Efficacy of Mathematics in Economics, *Handbook of the Philosophy of Science, Vol.5: Philosophy of Economics*, Elsevier, Oxford, pp. 159–198.

Mirowski, P. E. and Hands, W. D.: 1998, A Paradox Of Budgets: The Postwar Stabilization of American Neoclassical Demand Theory, *History of Political Economy* **30**((annual suppl.)), 260–292.

Mirowski, P. and Hands, W. D.: 2006, Introduction to Agreement on Demand: Consumer Theory in the Twentieth Century, *History of Political Economy* **38**(annual suppl.), 1–6.

Mongin, P.: 1998, Utility Theory and Ethics, *in* S. Barberà, P. J. Hammond and C. Seidle (eds), *Handbook of Utility Theory, Volume 1: Principles*, Kluwer Academic Publishers, Boston, pp. 371–481.

Mongin, P.: 2000, Les préférences révélées et la formation de la théorie du consommateur, *Revue économique* **51**(5), 1125–1152.

Mongin, P.: 2001, La théorie économique a-t-elle besoin des mathématiques?, *Commentaire* (93), 129–140.

Mongin, P.: 2003, L'axiomatisation et les théories économiques, *Revue Economique* **54**(1), 99–138.

Mongin, P.: 2004, "L'axiomatisation et les théories économiques": réponses aux critiques, *Revue Economique* **55**(1), 143–147.

Mongin, P.: 2006a, A Concept of Progress for Normative Economics, *Economics and Philosophy* **22**(01), 19–54.

Mongin, P.: 2006b, L'analytique et le synthétique en économie, *Recherches Economiques de Louvain* **72**(4), 349–383.

Mongin, P.: 2006c, Value Judgments and Value Neutrality in Economics, *Economica* **73**(290), 257–286.

Mongin, P.: 2009, Duhemian Themes in Expected Utility Theory, *in* A. Brenner and J. Gayon (eds), *French Studies in the Philosophy of Science*, Springer Netherlands, Dordrecht, pp. 303–357.

Mongin, P.: 2011, La theorie de la decision et la psychologie du sens commun, *Social Science Information* **50**(3-4), 351–374.

Mongin, P.: 2014, Le Paradoxe d'Allais: Comment lui rendre sa signification perdue? (Allais's Paradox: How to Give It Back Its Lost Meaning?), *Revue Economique* **65**(5), 743–779.

Mongin, P. and Pivato, M.: 2015, Ranking Multidimensional Alternatives and Uncertain Prospects, *Journal of Economic Theory* **157**(May), 146–171.

Monsell, S. and Driver, J.: 2000, Banishing the Control Homunculus, *in* S. Monsell and J. Driver (eds), *Control of Cognitive Processes: Attention and Performance XVIII*, The MIT Press, Cambridge, MA and London.

Moons, W. G., Mackie, D. M. and Garcia-Marques, T.: 2009, The Impact of Repetition-Induced Familiarity on Agreement With Weak and Strong Arguments, *Journal of Personality and Social Psychology* **96**(1), 32–44.

Morgan, M. S.: 2012, *The World in the Model: How Economists Work and Think*, Cambridge University Press, New York.

Moscati, I.: 2012, Intension, Extension and the Model of Belief and Knowledge in Economics, *Erasmus Journal for Philosophy and Economics* **5**(2), 1–26.

Moscati, I.: 2013a, How Cardinal Utility Entered Economic Analysis, 1909-1944, *European Journal of the History of Economic Thought* **20**(6), 906–939.

Moscati, I.: 2013b, Were Jevons, Menger and Walras Really Cardinalists? On the Notion of Measurement in Utility Theory, Psychology, Mathematics and Other Disciplines, 1870-1910, *History of Political Economy* **45**(3), 373–414.

Moscati, I.: 2016a, How Economists Came to Accept Expected Utility Theory: The Case of Samuelson and Savage, *Journal of Economic Perspectives* **30**(Forthcoming).

Moscati, I.: 2016b, *Measuring Utility: from the Marginal Revolution to Neuroeconomics*, forthcomin edn, Oxford University Press, Oxford.

Moxey, L. M.: 2011, Mechanisms Underlying Linguistic Framing Effects, *in* G. Keren (ed.), *Perspectives on Framing*, Psychology Press, New York, pp. 119–134.

Mullainathan, S. and Thaler, R. H.: 2001, Behavioral Economics, *in* N. J. Smelser and P. B. Baltes (eds), *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier Science Ltd., Amsterdam, pp. 1094–1100.

Mulligan, N. W. and Peterson, D. J.: 2013, The Negative Repetition Effect, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **39**(5), 1403–1416.

Nagatsu, M.: 2013, Experimental Philosophy of Economics, *Economics and Philosophy* **29**(2), 263–276.

Novakova, J. and Flegr, J.: 2013, How Much Is Our Fairness Worth? The Effect of Raising Stakes on Offers by Proposers and Minimum Acceptable Offers in Dictator and Ultimatum Games, *PLoS ONE* **8**(4), 1–9.

O'Donoghue, T. and Rabin, M.: 1999a, Doing It Now or Later, *The American Economic Review* **89**(1), 103–124.

O'Donoghue, T. and Rabin, M.: 1999b, Incentives for Procrastinators, *The Quarterly Journal of Economics* **114**(3), 769–816.

O'Donoghue, T. and Rabin, M.: 2001, Choice and Procrastination, *The Quarterly Journal of Economics* **116**(1), 121–160.

O'Donoghue, T. and Rabin, M.: 2003, Studying Optimal Paternalism, Illustrated by a Model of Sin Taxes, *The American Economic Review* **93**(2), 186–191.

O'Donoghue, T. and Rabin, M.: 2005, Optimal Taxes for Sin Goods, *Swedish Economic Policy* **12**(2), 7–39.

O'Donoghue, T. and Rabin, M.: 2006, Optimal sin taxes, *Journal of Public Economics* **90**(10-11), 1825–1849.

Ok, E. A., Ortoleva, P. and Riella, G.: 2015, Revealed (P)Reference Theory, *American Economic Review* **105**(1), 299–321.

Okder, H.: 2012, The Illusion of the Framing Effect in Risky Decision Making, *Journal of Behavioral Decision Making* **25**(1), 63–73.

Oppenheim and Putnam, H.: 1958, The Unity of Science as a Working Hypothesis, *in* H. Feigl, M. Scriven and G. Maxwell (eds), *Minnesota Studies in the Philosophy of Science, vol. 2*, Minnesota University Press, Minneapolis, pp. 3–35.

Peleg, B. and Yaari, M.: 1973, On the Existence of a Consistent Course of Action When Tastes are Changing, *The Review of Economic Studies* **40**(3), 391–401.

Pellé, S.: 2009, *Amartya K. Sen: La Possibilité d'une Ethique Economique Rationnelle*, PhD thesis, Université Paris 1 Panthéon-Sorbonne.

Percoco, M. and Nijkamp, P.: 2009, Estimating Individual Rates of Discount: A Meta-analysis., *Applied Economics Letters* **16**(12), 1235–1239.

Pesendorfer, W.: 2006, Behavioral Economics Comes of Age: A Review Essay on Advances in Behavioral Economics, *Journal of Economic Literature* **XLIV**(September), 712–721.

Pettit, P.: 1991, Decision Theory and Folk Psychology: Issues and Advances, *in* M. O. Bacharach and S. Hurley (eds), *Foundations of Decision Theory*, Blackwell, Oxford, pp. 147–175.

Phelps, E. S. and Pollak, R. A.: 1968, On Second-Best National Saving and Game-Equilibrium Growth, *The Review of Economic Studies* **35**(2), 185–199.

Plott, C. R. and Zeiler, K.: 2011, The Willingness to Pay –Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: Reply, *The American Economic Review* **101**(2), 1012–1028.

Pollak, R. A.: 1968, Consistent Planning, *The Review of Economic Studies* **35**(2), 201–208.

Pollak, R. A.: 1990, Distinguished Fellow: Houthakker's Contributions to Economics, *Journal of Economic Perspectives* **4**(2), 141–156.

Posner, R.: 1987, Charles Morris and the Behavioral Foundations of Semantics, *in* M. Krampen, K. Oehler, R. Posner, T. A. Sebeok and T. von Uexküll (eds), *Classics of Semiotics*, Springer, New York, pp. 23–58.

Pratt, J. W.: 1964, Risk Aversion in the Small and in the Large, *Econometrica* **32**(1), 122–136.

Prelec, D. and Loewenstein, G.: 1991, Decision Making Over Time and Under Uncertainty: A Common Approach, *Management Science* **37**(7), 770–786.

Prelec, D. and Loewenstein, G.: 1998, The Red and the Black: Mental Accounting of Savings and Debt, *Marketing Science* **17**(1), 4–28.

Putnam, H.: 1971, *Philosophy of Logic*, Routledge, New York.

Putnam, H.: 2002, *The Collapse of the Fact/Value Dichotomy and Other Essays*, Harvard University Press, Cambridge, MA and London.

Putnam, H.: 2004, *Ethics Without Ontology*, Harvard University Press, Cambridge, MA and London.

Putnam, H.: 2013, *Philosophy in an Age of Science*, Harvard University Press, Cambridge, England.

Putnam, H.: 2015, Naturalism, realism, and Normativity, *Journal of the American Philosophical Association* **1**(2), 312–328.

Putnam, H. and Walsh, V.: 2011, *The End of Value-Free Economics*, Routledge, Abingdon.

Quattrone, G. A. and Tversky, A.: 1988, Contrasting rational and psychological analyses of political choice, *The American Political Science Review* **82**(3), 719–736.

Quiggin, J.: 1982, A Theory of Anticipated Utility, *Journal of Economic Behavior & Organization* **3**(4), 323–343.

Quiggin, J.: 1993, *Generalized Expected Utility Theory. The rank-dependent model*, Springer, Dordrecht.

Quiggin, J.: 2014, Non-Expected Utility Models Under Objective Uncertainty, *in* M. J. Machina and K. W. Viscusi (eds), *Handbook of the Economics of Risk and Uncertainty*, Elsevier B.V., Oxford, pp. 701–728.

Quiggin, J. and Wakker, P. P.: 1994, The Axiomatic Basis of Anticipated Utility, *Journal of Economic Theory* **64**(2), 486–499.

Quine, W.: 1951, Two Dogmas of Empiricism, *The Philosophical Review* **60**(1), 20–43.

Quine, W. v. O.: 1994, Promoting Extensionality, *Synthese* **98**, 143–151.

Rabin, M.: 1993, Incorporating Fairness into Game Theory and Economics, *The American Economic Review* **83**(5), 1281–1302.

Rabin, M.: 1996, Psychology and Economics, *Working Paper* .
URL: *http://escholarship.org/uc/item/8jd5z5j2*

Rabin, M.: 1998, Psychology and Economics, *Journal of Economic Literature* **36**(1), 11–46.

Rabin, M.: 2000, Risk Aversion and Expected-Utility Theory: A Calibration Theorem, *Econometrica* **68**(5), 1281–1292.

Rabin, M.: 2002, A perspective on psychology and economics, *European Economic Review* **46**, 657 – 685.

Rabin, M.: 2013, An Approach to Incorporating Psychology into Economics, *The American Economic Review* **103**(3), 617–622.

Rabin, M. and Thaler, R. H.: 2001, Anomalies: Risk Aversion, *Journal of Economic Perspectives* **15**(1), 219–232.

Rastier, F.: 1996, Représentation ou interprétation?, *in* V. Rialle and D. Fisette (eds), *Penser l'esprit: des sciences de la cognition à une philosophie cognitive*, Presse Universitaire de Grenoble, Grenoble, pp. 219–232.

Rastier, F.: 2008, Le langage sans origine ou l'émergence du milieu sémiotique, *in* R. Delamotte-Legrand, C. Hudelot and A. Salaza Orvig (eds), *Dialogues, mouvements discursifs, significations: hommage à Frédéric François*, Cortil-Wodon: E.M.E., Fernelmont, pp. 207–222.

Rastier, F.: 2012, Langage et pensée : dualité sémiotique ou dualisme cognitif ?, *Texto !* **XVII**(1-2), 1–42.

Read, D., Loewenstein, G. and Rabin, M.: 1999, Choice Bracketing, *Journal of Risk and Uncertainty* **19**(1-3), 171–197.

Recanati, F.: 2006, Pragmatics and Semantics, *in* L. R. Horn and G. Ward (eds), *The Handbook of Pragmatics*, Blackwell, Malden, pp. 442–462.

Reyna, V. F., Chick, C. F., Corbin, J. C. and Hsia, A. N.: 2014, Developmental Reversals in Risky Decision Making: Intelligence Agents Show Larger Decision Biases Than College Students, *Psychological Science* **25**(1), 76–84.

Richter, M. K.: 1971, Rational Choice, *in* J. S. Chipman, L. Hurwicz, M. K. Richter and F. Sonnenschein Hugo (eds), *Preferences, Utility and Demand: A Minnesota Symposium*, Hartcourt Brace Jovanovich, New York, pp. 27–58.

Robinson, L. A. and Hammitt, J. K.: 2011a, Behavioral Economics and Regulatory Analysis, *Risk Analysis* **31**(9), 1408–22.

Robinson, L. A. and Hammitt, J. K.: 2011b, Behavioral Economics and the Conduct of Benefit-Cost Analysis: Towards Principles and Standards, *Journal of Benefit-Cost Analysis* **2**(2), 2–48.

Rosenberg, A.: 1992, *Economics – Mathematical Politics or Science of Diminishing Returns?*, The University of Chicago Press, Chicago and London.

Ross, D.: 2005, *Economic Theory and Cognitive Science: Microexplanation*, The MIT Press, Cambridge, MA and London.

Ross, D.: 2010, Economic Models of Procrastination, *in* C. Andreou and M. D. White (eds), *The Thief of Time: Philosophical Essays on Procrastination*, Oxford University Press, New York, pp. 28–50.

Ross, D.: 2014, *Philosophy of Economics*, Palgrave Macmillan, Basingstoke.

Ross, D., Ainslie, G. and Hofmeyr, A.: 2010, Self-control, Discounting and Rewards: Why picoeconomics is Economics, *Working Paper* .

Ross, D., Ladyman, J. and Spurett, D.: 2007, In Defence of Scientism, *in* J. Ladyman and D. Ross (eds), *Every Thing Must Go. Metaphysics Naturalized*, Oxford University Press, Oxford.

Rotemberg, J.: 2014, Models of Caring, or Acting As If One Cared, About the Welfare of Others, *Annual Review of Economics* **6**(1), 129–154.

Rothman, A. J. and Updegraff, J. A.: 2011, Specifying When and How Gain- and Loss-Framed Messages Motivates Healthy Behavior: An Integrated Approach, *in* G. Keren (ed.), *Perspectives on Framing*, Psychology Press, New York, pp. 257–278.

Rubinstein, A.: 2000, *Economics and Language. Five essays.*, Cambridge University Press, Cambridge.

Rubinstein, A.: 2012, *Economic Fables*, Open Book Publishers.
URL: *http://www.openbookpublishers.com/product/136/economic-fables*

Runciman, W. G. and Sen, A. K.: 1965, Games, Justice and the General Will, *Mind* **74**(296), 554–562.

Rustichini, A.: 2008, Dual or Unitary System? Two Alternative Models of Decision Making, *Cognitive, Affective, & Behavioral Neuroscience* **8**(4), 355–362.

Ryan, M. J.: 2005, Framing, Switching and Preference Reversals, *Theory and Decision* **57**(3), 181–211.

Sadock, J.: 2006, Speech Acts, *in* L. R. Horn and G. Ward (eds), *Handbook of Pragmatics*, Blackwell, Malden, pp. 53–73.

Saez, E.: 2007, Comment, *in* P. Diamond and H. Vartiainen (eds), *Behavioral Economics and its Applications*, Princeton University Press, Princeton and Oxford, pp. 81–84.

Sahlin, N.-e., Wallin, A. and Persson, J.: 2010, Decision Science: From Ramsey to Dual Process Theories, *Synthese* **172**(1), 129–143.

Saito, K.: 2011, Strotz Meets Allais: Diminishing Impatience and the Certainty Effect: Comment, *The American Economic Review* **101**(5), 2271–2275.

Saito, K.: 2013, Social Preferences Under Risk: Equality of Opportunity versus Equality of Outcome, *American Economic Review* **103**(7), 3084–3101.

Samuelson, L.: 2005, Economic Theory and Experimental Economics, *Journal of Economic Literature* **43**(1), 65–107.

Samuelson, P. A.: 1937, A Note on Measurement of Utility, *The Review of Economic Studies* **4**(2), 155–161.

Sandel, M. J.: 2000, What Money Can't Buy: the Moral Limits of Markets, *Tanner Lecture on Human Values* **21**, 87–122.

Savage, L. J.: 1972, *The Foundations of Statistics (Second Revised Edition)*, Dover publications, New York.

Savage, L. J.: 1987, Letter to Robert Aumann, 27 January 1971, *in* J. H. Dreze (ed.), *Essays on Economic Decisions under Uncertainty*, Cambridge University Press, Cambridge, England, pp. 78–81.

Schinkus, C.: 2010, The Importance of Communicative Rationality on Financial Markets, *Journal of Economic and Social Research* **12**(2), 119–144.

Schmidt, U.: 2003, Reference Dependence in Cumulative Prospect Theory, *Journal of Mathematical Psychology* **47**(2), 122–131.

Schmidt, U., Starmer, C. and Sugden, R.: 2008, Third-generation Prospect Theory, *Journal of Risk and Uncertainty* **36**(3), 203–223.

Schneider, W. and Shiffrin, R. M.: 1977, Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention, *Psychological Review* **84**(1), 1–66.

Schotter, A.: 2008, What's So Informative about Choice?, *in* A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics: A Handbook*, Oxford University Press, New York, pp. 70–94.

Schul, Y.: 2011, Alive or Not Dead: Implications for Framing from Research on Negations, *in* G. Keren (ed.), *Perspectives on Framing*, Psychology Press, New York, pp. 157–176.

Schulte-Mecklenbeck, M. and Kühberger, A.: 2014, Out of sight – out of mind? Information acquisition patterns in risky choice framing, *Polish Psychological Bulletin* **45**(1), 21–28.

Searle, J. R.: 1965, What is a Speech Act?, *in* M. Black (ed.), *Philosophy in America*, Allen and Unwin, London, pp. 221–239.

Searle, J. R.: 1969, *Speech Acts. An Essay in the Philosophy of Language*, Oxford University Press, Oxford.

Searle, J. R.: 1979, *Expression and Meaning. Studies in the Theory of Speech Acts*, Cambridge University Press, Cambridge, England.

Searle, J. R.: 2007, Grice on Meaning: 50 Years Later, *Teorema* **XXVI**(2), 9–18.

Searle, J. R. and Vanderveken, D.: 1985, *Foundations of Illocutionary Logic*, Cambridge University Press, Cambridge, England.

Sen, A. K.: 1961, On Optimising the Rate of Saving, *The Economic Journal* **71**(283), 479–496.

Sen, A. K.: 1971, Functions and Revealed Preference, *The Review of Economic Studies* **38**(3), 307–317.

Sen, A. K.: 1973, Behaviour and the Concept of Preference, *Economica* **40**(159), 241–259.

Sen, A. K.: 1977a, On Weights and Measures: Informational Constraints in Social Welfare Analysis, *Econometrica* **45**(7), 1539–1572.

Sen, A. K.: 1977b, Rational Fools: A Critique of the Behavioral Foundations of Economic Theory, *Philosophy and Public Affairs* **6**(4), 317–344.

Sen, A. K.: 1979, Informational Analysis of Moral Principles, *in* R. Harrison (ed.), *Rational Action: Studies in Philosophy and Social Sciences*, University of Cambridge, Cambridge, England, pp. 115–132.

Sen, A. K.: 1980, Description as Choice, *Oxford Economic Papers* **32**(3), 353–369.

Sen, A. K.: 1982, Rights and Agency, *Philosophy & Public Affairs* **11**(1), 3–39.

Sen, A. K.: 1983, Liberty and Social Choice, *The Journal of Philosophy1* **80**(1), 5–28.

Sen, A. K.: 1987, *On Ethics & Economics*, Blackwell, Malden.

Sen, A. K.: 1995, Is the Idea of Purely Internal Consistency of Choice Bizarre?, *in* J. Altham and R. Harrison (eds), *World, Mind, and Ethics. Essays on the Ethical Philosophy of Bernard Williams*, Cambridge University Press, Cambridge, England, pp. 19–31.

Sen, A. K.: 2002, *Rationality and Freedom*, Harvard University Press, Cambridge, MA.

Sen, A. K.: 2005, Walsh on Sen after Putnam, *Review of Political Economy* **17**(1), 107–113.

Sen, A. K.: 2006, *Identity and Violence: The Illusion of Destiny*, W. W. Norton & Company, New York.

Sen, A. K.: 2007, Why Exactly is Commitment Important for Rationality, *in* F. Peter and H. B. Schmid (eds), *Rationality and Commitment*, Oxford University Press, Oxford, pp. 17–27.

Sen, A. K.: 2009, *The Idea of Justice*, Harvard University Press, Cambridge, MA.

Sent, E.-M.: 2004, Behavioral Economics: How Psychology Made Its (Limited) Way Back Into Economics, *History of Political Economy* **36**(4), 735–760.

Shafir, E.: 1993, Choosing versus Rejecting: Why Some Options are Both Better and Worse than Others, *Memory & Cognition* **21**(4), 546–556.

Shafir, E.: 2013, *The Behavioral Foundations of Public Policy*, Princeton University Press, Princeton and Oxford.

Shafir, E., Diamond, P. and Tversky, A.: 1997, Money Illusion, *The Quarterly Journal of Economics* **112**(2), 341–374.

Shafir, E., Simonson, I. and Tversky, A.: 1993, Reason-based Choice, *Cognition* **49**(1-2), 11–36.

Shafir, E. and Thaler, R. H.: 2006, Invest Now, Drink Later, Spend Never: On the Mental Accounting of Delayed Consumption, *Journal of Economic Psychology* **27**(5), 694–712.

Shefrin, H. M. and Thaler, R. H.: 1988, The Behavioral Life-Cycle Hypothesis, *Economic Inquiry* **26**(4), 609–43.

Sher, S. and McKenzie, C. R. M.: 2006, Information Leakage from Logically Equivalent Frames, *Cognition* **101**(3), 467–94.

Sher, S. and McKenzie, C. R. M.: 2008, Framing Effects and Rationality, *in* N. Chater and M. Oaksford (eds), *The Probabilistic Mind: Projects for Bayesian Cognitive Science*, Oxford University Press, chapter 4, pp. 79–96.

Sher, S. and McKenzie, C. R. M.: 2011, Levels of Information: A Framing Hierarchy, *in* G. Keren (ed.), *Perpectives on Framing*, Psychology Press, New York, chapter 2, pp. 35–64.

Shiell, A. and Rush, B.: 2003, Can Willingness to Pay Capture the Value of Altruism? An Exploration of Sen's Notion of Commitment, *Journal of Socio-Economics* **32**(6), 647–660.

Shiffrin, R. M. and Schneider, W.: 1977, Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending and a General Theory, *Psychological Review* **84**(2), 127–190.

Shiv, B. and Fedorikhin, A.: 1999, Heart and Mind in Conflict: The Interplay of Affect and Cognition in Consumer Decision Making, *Journal of Consumer Research* **26**(3), 278–292.

Sieck, W. and Yates, F. J.: 1997, Exposition Effects on Decision Making: Choice and Confidence in Choice, *Organizational Behavior and Human Decision Processes* **70**(3), 207–219.

Simonson, I. and Tversky, A.: 1992, Choice in Context: Tradeoff Contrast and Extremeness Aversion, *Journal of Marketing Research* **XXIX**(August), 281–95.

Sinaceur, M., Heath, C. and Cole, S.: 2005, Emotional and Deliberative Reactions to a Public Crisis: Mad Cow Disease in France., *Psychological science* **16**(3), 247–54.

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

Smith, B.: 1990, Towards a History of Speech Act Theory, *in* Burkhardt (ed.), *Speech Acts, Meanings and Intentions. Critical Approaches to the Philosophy of John Searle*, de Gruyter, Berlin/New York, pp. 29–61.

Smith, V. L.: 1985, Experimental Economics: Reply, *The American Economic Review* **75**(1), 265–272.

Sobel, J.: 2005, Interdependent Preferences and Reciprocity, *Journal of Economic Literature* **XLIII**(2), 392–436.

Spiegler, R.: 2013, Comments on "Behavioral" Decision Theory, *in* D. Acemoglu, M. Arellano and E. Dekel (eds), *Advances in Economics and Econometrics: Tenth World Congress, Volume 1*, Cambridge University Press, New York, pp. 289–303.

Stanovich, K. E. and West, R. F.: 2000, Individual Differences in Reasoning: Implications for the Rationality Debate?, *The Behavioral and Brain Sciences* **23**(5), 645–65.

Starmer, C.: 2000, Developments in Non-Expected Utility Theory: the Hunt for a Descriptive Theory of Choice Under Risk, *Journal of Economic Literature* **XXXVIII**(2), 332–382.

Starmer, C.: 2005, Normative Notions in Descriptive Dialogues, *Journal of Economic Methodology* **12**(2), 277–289.

Stein, E.: 1996, *Without Good Reasons: The Rationality Debate in Philosophy and Cognitive Science*, Clarendon Press, Oxford.

Stigler, G. and Becker, G.: 1977, De Gustibus Non Est Disputandum, *The American Economic Review* **2**(67), 76–90.

Stommel, E.: 2013, *Reference-Dependent Preferences: A Theoretical and Experimental Investigation of Individual Reference-Point Formation*, Springer.

Strawson, G.: 2004, Against Narrativity, *Ratio* **17**(4), 428–452.

Strotz, R. H.: 1956, Myopia and Inconsistency in Dynamic Utility Maximization, *The Review of Economic Studies* **23**(3), 165–180.

Su, H.-C. and Colander, D.: 2013, A Failure to Communicate: The Fact-value Divide and the Putnam-Dasgupta Debate, *Erasmus Journal for Philosophy and Economics* **6**(2), 1–23.

Sugden, R.: 1991, Rational Choice: A Survey of Contributions from Economics and Philosophy, *The Economic Journal* **101**(407), 751–785.

Takemura, K.: 2014, *Behavioral decision theory. Psychological and mathematical descriptions of human choice behavior*, Springer, Tokyo.

Taleb, N. N.: 2007, *The Black Swan: The Impact of the Highly Improbable*, Random House, New York.

Tanaka, T., Camerer, C. F. and Nguyen, Q.: 2010, Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam, *The American Economic Review* **100**(1), 557–571.

Tarski, A.: 2009, Du concept de conséquence logique, *in* D. Bonnay and M. Cozic (eds), *Textes clés de philosophie de la logique*, Vrin, Paris, pp. 83–98.

Teigen, K. H.: 2011, When Frames Meet Realities: On the Perceived Correctness of Inaccurate Estimates, *in* G. Keren (ed.), *Perpectives on Framing*, Psychology Press, New York, pp. 197–218.

Thaler, R. H.: 1980, Toward a Positive Theory of Consumer Choice, *Journal of Economic Behavior & Organization* **1**(1), 7–59.

Thaler, R. H.: 1981, Some Empirical Evidence on Dynamic Inconsistencies, *Economics Letters* **8**(3), 201–207.

Thaler, R. H.: 1985, Mental Accounting and Consumer Choice, *Marketing Science* **4**(3), 199–214.

Thaler, R. H.: 1987, The Psychology of Choice and the Assumptions of Economics, *in* A. Roth (ed.), *Laboratory Experimentation in Economics. Six points of view*, Cambridge University Press, Cambridge, MA.

Thaler, R. H.: 1992, *The Winner's Curse: Paradoxes and Anomalies of Economic Life*, Princeton University Press, Princeton.

Thaler, R. H.: 1999, Mental Accounting Matters, *Journal of Behavioral Decision Making* **12**(3), 183–206.

Thaler, R. H.: 2000, From Homo Economicus to Homo Sapiens, *Journal of Economic Perspectives* **14**(1), 133–141.

Thaler, R. H.: 2008, Commentary – Mental Accounting and Consumer Choice: Anatomy of a Failure, *Marketing Science* **27**(1), 15–25.

Thaler, R. H.: 2015, *Misbehaving: the Making of Behavioral Economics*, W. W. Norton & Company, New York.

Thaler, R. H.: 2016, AEA Presidential Address: Behavioral Economics: Past, Present and Future, *American Economic Review* (Forthcoming).
URL: *https://www.aeaweb.org/webcasts/2016/Behavioral.php*

Thaler, R. H. and Johnson, E. J.: 1990, Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice, *Management Science* **36**(6), 643–660.

Thaler, R. H. and Shefrin, H. M.: 1981, An Economic Theory of Self-Control, *The Journal of Political Economy* **89**(2), 392–406.

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

Thaler, R. H. and Sunstein, C. R.: 2008, *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Yale University Press, New Haven & London.

Thaler, R. H., Sunstein, C. R. and Balz, J. P.: 2013, Choice Architecture, *in* E. Shafir (ed.), *The Behavioral Foundations of Public Policy*, Princeton University Press, Princeton and Oxford, pp. 428–439.

Thiel, C.: 2003, The Extension of the Concept Abolished? Reflexions on a Fregean Dilemma, *in* J. Hintikka, T. Czarnecki, K. Kijania-Placek, T. Placek and A. Rojszczak (eds), *Philosophy and Logic: In Search of the Polish Tradition*, Kluwer Academic Publishers, pp. 269–274.

Thiel, C.: 2009, Gottlob Frege and the Interplay between Logic and Mathematics, *in* L. Haaparanta (ed.), *The Development of Modern Logic*, Oxford University Press, New York, pp. 196–202.

Tombu, M. and Mandel, D. R.: 2015, When Does Framing Influence Preferences, Risk Perceptions, and Risk Attitudes? The Explicated Valence Account, *Journal of Behavioral Decision Making* **28**(5), 464–476.

Toplak, M. E., West, R. F. and Stanovich, K. E.: 2014, Assessing Miserly Information Processing: An Expansion of the Cognitive Reflection Test, *Thinking & Reasoning* **20**(2), 147–168.

Trautmann, S. T. and Vieider, F. M.: 2012, Social Influences on Risk Attitudes: Applications in Economics, *in* S. Roeser, R. Hillerbrand, P. Sandin and M. Peterson (eds), *Handbook of Risk Theory: Epistemological, Decision Theory, Ethics, and Social Implications of Risk*, Springer, Dordrecht, pp. 575–600.

Trautmann, S. T. and Wakker, P. P.: 2010, Process Fairness and Dynamic Consistency, *Economics Letters* **109**(3), 187–189.

Turner, S. P.: 2010, *Explaining the Normative*, Polity, Cambridge, England.

Tversky, A.: 1975, A Critique of Expected Utility Theory: Descriptive and Normative Considerations, *Erkenntnis* **9**(2), 163–173.

Tversky, A.: 1996, Rational Theory and Constructive Choice, *in* K. J. Arrow, E. Colombatto, M. Perlman and C. Schmidt (eds), *The Rational Foundations of Economic Behavior*, MacMillan, Basingstoke, pp. 185–197.

Tversky, A. and Kahneman, D.: 1981, The Framing of Decisions and the Psychology of Choice, *Science* **211**(4481), 453–458.

Tversky, A. and Kahneman, D.: 1986, Rational Choice and the Framing of Decisions, *The Journal of Business* **59**(4), S251–S278.

Tversky, A. and Kahneman, D.: 1991, Loss Aversion in Riskless Choice: A Reference-Dependent Model, *Quarterly Journal of Economics* **106**(4), 1039–1061.

Tversky, A. and Kahneman, D.: 1992, Advances in Prospect Theory: Cumulative Representation of Uncertainty, *Journal of Risk and Uncertainty* **5**(4), 297–323.

Tversky, A. and Koehler, D. J.: 1994, Support Theory: A Nonextensional Representation of Subjective Probability, *Psychological Review* **101**(4), 547–567.

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

Vallois, N.: 2011, The Pathological Paradigm of Neuroeconomics, *Oeconomia* **1**(4), 525–556.

Vallois, N.: 2012a, Des décideurs incohérents mais rationnels: la réhabilitation normative de l'incohérence séquentielle par la psychologie évolutionniste, *Working Paper* .
URL: *https://paris1.academia.edu/NicolasVallois*

Vallois, N.: 2012b, *L'émergence d'un nouveau domaine de savoir: la neuroéconomie*, Ph.d, Paris 1 Panthéon-Sorbonne.

Vallois, N.: 2014, Neurosciences et politiques publiques : Vers un nouvel interventionnisme économique, *Revue de Philosophie Economique* **15**(2), 132–175.

van Buiten, M. and Keren, G.: 2009, Speaker-listener Incompatibility: Joint and Separate Processing in Risky Choice Framing, *Organizational Behavior and Human Decision Processes* **108**(1), 106–115.

Varian, H. R.: 2005, *Intermediate Microeconomics 7th edition*, W. W. Norton & Company.

Vilks, A.: 1995, On Mathematics and Mathematical Economics, *Greek Economic Review* **17**(2), 177–204.

Vilks, A.: 1998, Axiomatization, *in* J. B. Davis, W. Hands and U. Mäki (eds), *The Handbook of Economic Methodology*, Edward Elgar Publishing, Cheltenham, pp. 25–28.

Von Neuman, J. and Morgenstern, O.: 1953, *Theory of Games and Economic Behavior*, 3 edn, Princeton University Press, Princeton.

Vromen, J.: 2011, Neuroeconomics: Two Camps Gradually Converging: What Can Economics Gain From it?, *International Review of Economics* **58**(3), 267–285.

Wakker, P. P.: 1994, Separating Marginal Utility and Probabilistic Risk Aversion, *Theory and Decision* **36**(1), 1–44.

Wakker, P. P.: 2010, *Prospect Theory for Risk and Ambiguity*, Cambridge University Press, Cambridge, England.

Wakker, P. P. and Tversky, A.: 1993, An Axiomatization of Cumulative Prospect Theory, *Journal of Risk and Uncertainty* **7**(2), 147–175.

Walsh, V.: 1996, Rationality, Allocation, and Reproduction, *Technical report*, Oxford.

Wänke, M. and Reutner, L.: 2011, Direction-of-Comparison Effects: How and Why Comparing Apples with Oranges is Different from Comparing Oranges with Apples, *in* G. Keren (ed.), *Perpectives on Framing*, Psychology Press, New York, pp. 177–194.

Weber, F.: 1996, Réduire ses dépenses, ne pas compter son temps. Comment mesurer l'économie domestique?, *Genèses* **25**(déc.), 5–28.

Weintraub, R. E.: 1998, Controversy: Axiomatisches Missverständnis, *The Economic Journal* **108**(451), 1837–1847.

Weintraub, R. E.: 2002, *How Economics Became a Mathematical Science*, Duke University Press, Durham and London.

Wilkinson, N. and Klaes, M.: 2012, *An Introduction to Behavioral Economics. 2nd Edition*, Palgrave Macmillan, Basingstoke.

Yi, R., Charlton, S., Porter, C., Carter, A. E. and Bickel, W. K.: 2011, Future Altruism: Social Discounting of Delayed Rewards, *Behavioural Processes* **86**(1), 160–163.

Yu, R. and Zhang, P.: 2014, Neural Evidence for Description Dependent Reward Processing in the Framing Effect, *Frontiers in Neuroscience* **8**(56), 1–11.

Zeckhauser, R.: 2014, Preface 2, *in* M. J. Machina and W. K. Viscusi (eds), *Handbook of the Economics of Risk and Uncertainty*, Elsevier and North-Holland, Amsterdam, pp. xvii–xxix.

Zhang, W. and Grenier, G.: 2013, How can Language be Linked to Economics? A survey of Two Strands of Research, *Language Problems and Language Planning* **37**(3), 203–226.

Zizzo, D. J.: 2010, Experimenter Demand effects in Economic Experiments, *Experimental Economics* **13**(1), 75–98.

Origins of the chapters cited from Sen (2002)

Chap.1: "Introduction : Rationality and Freedom", pp.3-65; previously unpublished.

Chap.3: "Internal Consistency of Choice", pp.121-157; originally published in *Econometrica*, 1993, 61(3), pp.495-521.

Chap.5: "Goals, Commitment, and Identity", pp.206-224; originally published in the *Journal of Law, Economics, and Organization*, 1985, 1(2), pp.341-355.

Chap.6: "Rationality and Uncertainty", pp.225-244; originally published in *Recent Developments in the Foundations of Utility and Risk Theory* (pp.3-25), in 1986, by Reidl, at Dodrecht, and edited by Daboni, L., Montesano, A., and Lines, M..

Chap.11: "Information and Invariance in Normative Choice", pp.349-380; originally published in *Social Choice and Public Decision Making: Essays in Honor of Kenneth J. Arrow, vol.1*, in 1986, by Cambridge University Presss, at Cambridge England, and edited by Heller W.P., Starr R., and Starrett D.A..

Traductions[21]

Introduction générale

« Le langage ordinaire et la pensée ordinaire ont été, en règle générale, victimes de négligence dans la science économique. Leur potentiel pour l'explication des phénomènes les plus économiques peut, toutefois, être grand. » (Bacharach, 1990, p.368)

« Les économistes semblent avoir largement ignoré le langage. C'est une situation malheureuse. Le monde dans lequel nous vivons est un monde de mots, pas de fonctions, et de nombreux phénomènes réels pourraient être plus facilement analysés si nous prenions cela en compte. » (Lipman 2003, p.75)

« La plus grande partie du partage de l'information n'est pas accomplie par du signalement à la Spence, par l'intermédiaire du système des prix, ni au travers de mécanismes d'incitations compatibles à la Hurwicz : il se fait par la discussion informelle, ordinaire. » (Farell et Rabin 1996, p.104)

Au cours des trois dernières décennies, l'analyse économique des comportements individuels a été soumise aux critiques incessantes d'une approche connue sous le nom d' « économie com-

[21] Nous avons également traduit par nos soins les citations.

portementale ». Un ensemble croissant de régularités empiriques est présenté par les économistes comportementaux comme contredisant les prévisions des modèles théoriques utilisés en économie dite « standard ». Ces critiques sont accompagnées de considérations méthodologiques diverses à propos, notamment, de l'absence de relations interdisciplinaires entre l'économie et la psychologie, ou de la distinction positif/normatif dans les modèles de comportements individuels. Au cours de la dernière décennie, les économistes qui se disaient ou se disent toujours « standards » ont de plus en plus répondu à ces critiques. Ces réponses ne sont pas homogènes. Certains économistes standards ont accepté une partie de ces critiques ; d'autres en ont accepté l'entièreté, et pratiquent désormais l'économie comportementale en se disant « économistes comportementaux ». Néanmoins, un ensemble d'économistes standards ont retourné la critique, arguant notamment que les économistes comportementaux ne comprennent pas correctement l'approche standard. En bref, la frontière entre l'économie standard et l'économie comportementale est plus floue aujourd'hui qu'elle ne l'était il y a trente ans. Ce faisant, les problèmes théoriques, empiriques et méthodologiques constituant cette frontière ont des connexions moins claires qu'auparavant. Cette frontière floue entre l'économie comportementale et l'économie standard est l'objet de cette thèse. Dans sa forme la plus générale, l'objectif est de fournir une meilleure compréhension des transformations théoriques et méthodologiques, impulsées par l'économie comportementale, que traverse aujourd'hui l'analyse économique des comportements individuels.

Si cette frontière est floue, c'est en partie liée à la stratégie de l'économie comportementale : s'inspirer de la psychologie pour *modifier* les modèles standards de comportements individuels. En d'autres termes, l' *homo œconomicus* traditionnel – dont le raisonnement est toujours correct, les choix effectués dans le passé ne sont jamais regrettés et guidés seulement par l'intérêt propre – est modifié pour prendre un visage plus humain, selon les économistes comportementaux. Cet argument est systématisé par le père de l'économie comportementale, Richard Thaler (2015; 2016), qui met en contraste deux espèces, les « Econs » représentés dans les modèles d'économie standard et les « Humains » représentés dans son approche. Dans l'économie comportementale, soutient-il, les agents économiques sont représentés comme commettant *parfois* des erreurs dans leurs raisonnements et dans leurs choix, en plus de ne *pas toujours* se comporter en regard de leurs intérêts propres. Ainsi l'économie comportementale est présentée comme une généralisation de

l'économie standard : les Econs ne se comportent jamais comme des humains, mais les humains se comportent parfois comme des Econs. L'une des raisons pour lesquelles Thaler et d'autres économistes comportementaux ne critiquent pas complètement l'économie standard est qu'ils en ont besoin pour caractériser ce qui compte comme comportements *rationnels* : l'économie standard est une référence normative pour l'économie comportementale. Par conséquent, la psychologie est utilisée en économie comportementale dans le but de modéliser les écarts entre la façon dont les agents économiques *se comportent* et comment ils *doivent* se comporter de façon rationnelle. L'objectif de cette thèse est d'essayer de montrer qu'un aspect non trivial de ce qui rend les agents économiques humains reste en grande partie inexploré dans l'économie comportementale, et que l'exploration de celui-ci permet d'éclairer les débats entre économie comportementale et économie standard. Cet aspect est tout simplement que nous parlons. En effet, les manières dont nous utilisons notre langage ne sont pas partagées par les autres espèces, y compris par les Econs. Comment les usages de langage jouent-ils un rôle dans la rationalité économique qui pourrait nous aider à mieux comprendre les débats entre économie comportementale et économie standard ?

Cette thèse soutient que nos usages de langage ne jouent pas un, mais deux rôles dans la rationalité économique. D'une part, les économistes utilisent du langage pour discuter et débattre sur la rationalité économique, ainsi que pour la théoriser. D'autre part, les agents économiques utilisent du langage pour faire des choix économiques ou, plus important encore, pour construire des problèmes de décisions dans lesquels d'autres agents font des choix économiques. Parmi le petit nombre d'économistes qui se sont intéressés à la relation entre langage et rationalité économique, Ariel Rubinstein exprime ces deux rôles de manière particulièrement frappante :

> « *Les agents économiques* sont des êtres humains pour qui le langage est un outil central dans le processus de prise de décisions et de formation des jugements. Et [...] les autres "joueurs" importants dans la théorie économique - à savoir nous-mêmes, les *économistes théoriciens* – utilisent des modèles formels, mais ce ne sont pas simplement des modèles mathématiques; leurs significations découlent de leurs interprétations, qui sont exprimées en utilisant le langage quotidien. » (2000 pp.3-4, nos italiques)

En d'autres termes, le rôle du langage dans la rationalité économique est double : les comportements des agents économiques sont en partie constitués d'usages de langage, et il en est de même pour les activités professionnelles des économistes, particulièrement les théoriciens. Deux

ensembles de distinctions classiques utilisées par les linguistes et les philosophes du langage sont pertinents pour davantage clarifier comment le rôle du langage, dans la rationalité économique, est étudié dans cette thèse. Le premier ensemble, proposé par le linguiste et sémioticien Ferdinand de Saussure (1916, chap.3), distingue le langage, la langue et la parole. Brièvement, le langage réfère de manière abstraite à l'ensemble de tous les *moyens de communication* (les symboles, les gestes, les phonèmes, etc.), dont une articulation conventionnelle particulière constitue une langue utilisée dans une *communauté* (l'Anglais, le Français, un dialecte local, etc.), qui s'instancie *individuellement* dans la parole (des mots et phrases, écrits ou parlés). Le second ensemble, proposé par le philosophe et sémioticien Charles W. Morris en 1938, distingue trois types de relations entre les symboles constitutifs d'une langue (voir Posner 1987, Sect.1) : la syntaxe, la sémantique et la pragmatique. Brièvement, la syntaxe désigne les relations *entre symboles*, la sémantique désigne les relations *entre des symboles et d'autres entités* (des « objets », « choses », « personnes », etc.) qui ne sont ni des symboles ni l'humain qui les utilise, et la pragmatique désigne les relations entre *des symboles et l'humain qui les utilise* (quelles sont ses intentions? Qui est son interlocuteur? etc.). Dans cette thèse, notre attention se portera sur la parole et la pragmatique, des économistes participant au débat entre économie comportementale et économie standard d'une part, et des agents économiques dans les situations de prise de décision qui sont abordées dans ces débats (par exemple, dans des expériences de laboratoire, répondant à des enquêtes, effectuant une transaction sur un marché, etc.), d'autre part. Notre argument est que l'examen de ces utilisations de langage peut contribuer à donner une image plus claire des principales questions sous-jacentes aux débats entre économie comportementale et économie standard.[22]

Cette introduction générale présente trois de ces principales questions, qui sont abordées tout au long de la thèse, à savoir la question de l'unification théorique (0.1), la question de

[22]Pour une perspective historique et philosophique approfondie sur les distinctions présentées ici et d'autres utilisées par les linguistes et les philosophes du langage, voir Rastier (2012). Mongin (2006b;c) examine d'autres distinctions classiques vis-à-vis de la méthodologie économique, par exemple entre les *phrases* (entités linguistiques), les *énoncés* (significations de phrases en termes de sens et de référence), les *propositions* (significations de phrases en termes de vérité et fausseté) et les *énonciations* (actions de prononcer ou d'écrire des phrases par un locuteur); ou entre l'*utilisation* de mots et de phrases et la simple *mention* de mots et de phrases. D'autres économistes ont déjà soutenu que les usages du langage jouent un rôle spécifique dans la rationalité économique qui reste à être pleinement examiné (pour une vision d'ensemble, voir Zhang et Grenier 2013; Alcouffe 2013). Toutefois, la plupart (sinon la totalité) de ces contributions mettent l'accent sur une langue et non sur la parole. La relation entre ces contributions et cette thèse ne sera pas abordée ici, mais plutôt dans des travaux futurs, principalement parce qu'elles n'ont aucun lien avec l'économie comportementale (voir toutefois Chen 2013).

l'interdisciplinarité entre l'économie et la psychologie (0.2) et la distinction positif/normatif (0.3) ; avant de présenter la manière dont les contributions de la thèse sont articulées en chapitres (0.4).

0.1 L'unification théorique et les trois dimensions

Dans les termes de la distinction bien pratique établie par Esther-Mirjam Sent (2004) entre la « nouvelle » et l' « ancienne » économie comportementale, cette thèse porte strictement sur la nouvelle économie comportementale. L'ancienne économie comportementale découle des travaux d'Herbert Simon sur la rationalité limitée à partir du milieu des années 1950. La nouvelle économie comportementale (simplement « économie comportementale » désormais) découle de l'influence des psychologues Daniel Kahneman et Amos Tversky sur Thaler à partir de la fin des années 1970.[23]

Depuis cette date, et surtout en empruntant les méthodes expérimentales des psychologues, les économistes comportementaux ont étudié des régularités empiriques dans les comportements individuels qui s'écartent systématiquement et significativement des prévisions faites par les modèles standards. Diverses explications théoriques émanant de différents modèles étaient données à cette multitude de régularités comportementales (en caricaturant, il y avait un modèle par régularité). Cependant, des explications théoriques unificatrices sont de plus en plus recherchées depuis la moitié des années 2000 (un modèle pour plusieurs régularités). La plupart des modèles de l'économie comportementale avant cette date sont reproduits dans le volume *Advances in Behavioral Economics* (désormais *Advances*), édité par Colin Camerer, George Loewenstein et Matthew Rabin en 2004. La structure d'une partie de ce volume, intitulée « sujet de base », montre assez bien ce que les économistes comportementaux prennent comme étant les domaines les plus élémentaires de l'analyse économique. Dans cette partie, les contributions sont organisées en cinq sections, dont les deuxième, troisième et quatrième dont sont respectivement intitulées « préférences sur les résultats risqués et incertains », « choix intertemporels » et « équité et préférences sociales ». Cette façon par laquelle trois dimensions des comportements

[23] Les liaisons et contrastes entre « nouvelle » et « ancienne » économie comportementale ne sont pas abordés dans cette thèse, étant donné qu'ils ont déjà été discutés longuement ailleurs (voir, par exemple, Sent 2004; Klaes et Sent 2005, surtout p.47; Egidi 2012; Kao et Velupillai 2015; Geiger 2016).

économiques – 'l'incertitude', 'le temps' et 'autrui' – sont clairement *séparées* est omniprésente dans la littérature autour de l'économie comportementale.[24]

Cette séparation est également présente dans des ouvrages classiques de l'économie standard (par exemple, Deaton and Muellbauer, 1980; Mas-Colell et et al. 1995), quoique d'une façon moins saillante. Cette séparation est extrêmement explicite dans les écrits contemporains des économistes standards sur l'économie comportementale. Ce n'est pas seulement le cas dans les deux recensions-essais d'*Avances* par Drew Fudenberg (2006) et Wolfgang Pesendorfer (2006), mais aussi, par exemple, dans la critique de l'économie comportementale par David Levine (2012). Concernant l'unification théorique, il convient de noter la position de Fudenberg:

> « Les économistes comportementaux (*et les économistes théoriciens!*) devraient consacrer plus d'effort à synthétiser les modèles existants et à en développer d'autres, plus généraux, et moins d'effort à modéliser encore une autre observation comportementale particulière » (Fudenberg 2006, p.699, nos italiques)

Ce que Fudenberg demande c'est en réalité ce qu'il est en train de faire avec Levine (2006; 2011; 2012c), c'est-à-dire un modèle fournissant, pour les régularités observées par les économistes comportementaux, une explication unifiée dans les trois dimensions ensemble. En résumé, les relations théoriques et empiriques entre les trois dimensions des comportements économiques sont actuellement un domaine de l'analyse économique où les limites entre l'économie comportementale et l'économie standard sont particulièrement floues. L'une des contributions que la thèse tentera d'apporter est une étude des conditions sous lesquelles une unification théorique (un modèle pour plusieurs régularités) articulant les trois dimensions ensemble est possible. Pour ce faire, nous examinerons les usages de langage que font les économistes dans leurs modèles de comportements individuels dans les trois dimensions (par exemple, la notion de « cohérence dynamique ») et par les agents économiques dans les situations où les régularités comportementales sont observées (par exemple, marquer une distinction linguistique entre « maintenant » et « plus tard »).

Dans cette thèse, l'économie standard réfère à l'économie pratiquée et défendue par les économistes qui s'identifient eux-mêmes comme « standards » (par exemple, Fudenberg, Levine,

[24]Les deux autres sections considérées comme sujet de base par ces économistes comportementaux sont intitulées « dépendance à la référence et aversion aux pertes » et « théorie des jeux ». La première comporte des contributions sur les comportements individuels en situation de certitude qui auront une place spécifique dans cette thèse. Par contraste, la seconde comporte des contributions concernant les interactions stratégiques (entre au moins deux personnes) qui ne seront pas abordées dans cette thèse.

Pesendorfer) et qui ont débattu ou commenté de manière critique sur les contributions des économistes qui s'identifient eux-mêmes comme « comportementaux » (par exemple, Thaler, Camerer, Loewenstein, Rabin). Cependant, ces critiques de l'économie comportementale ont des positions différentes concernant ce que sont les principes fondateurs de l'économie standard. D'une part, les théoriciens des jeux, tels que Levine, soulignent que « le cœur de la théorie économique moderne « rationnelle » est le concept d'un équilibre non coopératif ou « Nash » d'un jeu » (Levine 2012, chap.2). Ce qu'est un équilibre de Nash peut être énoncé comme l'*identification* des actions rationnelles à prendre par (au moins) deux agents indépendamment l'un de l'autre, mais les conséquences respectives de ces actions pour chaque agent sont mutuellement dépendantes des actions choisies. Cette thèse ne sera pas concernée par ces questions de théories des jeux. Elle sera en revanche concentrée sur les questions que soulèvent les théoriciens de la décision tel que Pesendorfer, pour qui le principe fondateur de l'économie est la « méthodologie » ou le « principe » de la préférence révélée – qui sont les mots utilisés à la place de « théorie » (voir Pesendorfer Gul et 2008; voir aussi Dekel et Lipman 2010). Pour le propos de cette introduction générale, ce qu'est la préférence révélée peut être énoncé comme l'identification des choix effectués par les agents économiques (directement observables dans les données économiques) avec leurs préférences (indirectement observables dans l'esprit des agents).[25]

Par exemple, Pesendorfer présente ses principales préoccupations méthodologiques avec l'économie comportementale comme suit :

> « Les modèles de choix standards et les théories de l'économie comportementale diffèrent dans leur orientation lors de l'analyse de phénomènes « comportementaux » [...]. Le travail théorique en économie comportementale met souvent l'accent sur la psychologie de la situation à l'aide du comportement économique comme une fenêtre sur l'esprit du preneur de décision [...]. Remarquez l'inversion des rôles de l'économie et de la psychologie. Les données économiques sont utilisées pour préciser les détails d'un processus mental particulier qui est opérationnel pour ce cas particulier. » (Pesendorfer 2006, pp.718-719)

[25] Une vision raffinée des relations entre l'économie « comportementale », « standard », « mainstream » et « néoclassique » est offerte par Colander (2000) et Davis (2006, 2007b, 2009b, 2011). Sur l'évolution historique et l'état contemporain des relations entre la théorie de la décision et la théorie des jeux, voir, par exemple, Mirowski (2002, 2006, 2009) ou Drèze et Aumann (2009). Sur comment cette relation s'inscrit dans l'évolution de la théorie de l'équilibre général, la théorie du choix du consommateur et la théorie de la demande, voir, par exemple, Kirman (1989) ou Mirowski et Hands (2006). Sur le statut contemporain de la préférence révélée, voir Hands (2012a; 2013b).

En d'autres termes, selon une perspective de théorie de la décision, le principal problème méthodologique avec l'économie comportementale est sa relation avec la psychologie; cette question est présentée dans la section suivante.

0.2 L'économie et la psychologie : le problème de l'interdisciplinarité

Pesendorfer est le coauteur avec Faruk Gul (2008) de l'une des critiques de l'économie comportementale les plus sévères et les plus discutées dans les années 2000, jusqu'à avoir déclenché un volume complet de réponses et de commentaires : *The Foundations of Positive and Normative Economics* (dorénavant *The Foundations*; Caplin et Schotter 2008a). Dans ce volume, les échanges critiques entre Gul et Pesendorfer (2008) *versus* Camerer (2008) donnent une illustration explicite du rôle des usages de langage, par les économistes, dans la rationalité économique par rapport au problème de l'interdisciplinarité entre économie et psychologie. À un certain moment de leur défense de l'économie standard, Gul et Pesendorfer mettent deux citations - une provenant de la psychologie et une provenant de l'économie – à côté l'une de l'autre, puis examinent le contraste entre elles (2008, p.5, références omises, nos italiques) :

> « Une bonne partie de l'aversion au risque est provoquée par des réponses de peur immédiate, qui sont en grande partie traçables à une petite zone du cerveau appelée l'amygdale.
> Un preneur de décision est (globalement) averse au risque, . . . si et seulement si son utilité von Neumann-Morgenstern est concave au niveau de richesse pertinent (à tous les niveaux de richesse).
>
> [. . .]La plupart des chercheurs reconnaissent les divers termes dans la deuxième citation comme des abstractions appartenant au vocabulaire spécialisé de l'économie. *Bien que moins apparent*, le langage dans la première citation est également spécialisé dans son utilisation d'abstraction spécifique à une discipline. Les termes « peur immédiate » et « traçables » sont des abstractions de la psychologie et des neurosciences. »

On notera que le point de Gul et Pesendorfer n'est ici pas seulement que les économistes et les psychologues utilisent un langage différent. C'est également que celui de l'économie (par exemple, « utilité von Neumann-Morgenstern ») est un *langage* plus *technique* (ou plus « spécifique à une discipline ») que celui de la psychologie (par exemple, « crainte immédiate ») qui est plus proche de notre *langage quotidien*. Dans sa réponse, Camerer fait valoir à propos de certaines régularités comportementales que :

> « Bon nombre de ces phénomènes peuvent être traduits dans le langage conventionnel de l'économie. En effet, une contribution majeure de la neuroéconomie [qui signifie également « économie comportementale » dans ce contexte - DJ] peut-être de fournir un *langage formel* pour parler de ces phénomènes. [...] D'autres phénomènes sont plus clairement compris par l'adoption de nouveaux termes provenant de la psychologie et des neurosciences plutôt que de se démener à intégrer l'activité complexe du cerveau maladroitement dans le lit de Procruste du langage économique. » (Camerer 2008, p.50, nos italiques)

Le point de Camerer est que le *langage formel* des mathématiques, de la logique et de la théorie des probabilités utilisé en économie est d'intérêt à la fois pour les économistes, mais aussi pour les psychologues (c'est même « une contribution majeure » de l'économie comportementale); mais l'utilisation de ce langage formel ne devrait pas empêcher les économistes, lorsqu'ils en ressentent le besoin, de l'interpréter avec le langage utilisé par les psychologues.[26]

Une contribution proposée dans cette thèse est d'étoffer de manière systématique cette perspective sur les utilisations du langage des économistes et des psychologues sur le problème de l'interdisciplinarité entre économie et psychologie. Pour ce faire, nous utiliserons les catégories de langage *formel*, *technique* et *quotidien* comme suit. Le langage formel désigne le symbolisme et la syntaxe des mathématiques, de la logique et de la théorie des probabilités tel que les économistes les utilisent pour construire par exemple les « modèles formels » dont parlait Rubinstein – eux-mêmes utilisés à diverses fins telles que la mesure de certaines variables, l'inférence de certaines conclusions, etc. (voir Boumans 2005; Morgan 2012). Le langage technique réfère à une langue qui est utilisée de manière « spécifique à une discipline » au sein d'une communauté scientifique. Et le langage quotidien réfère à nos usages de langage dans nos vies quotidiennes qui n'ont *a priori* rien à voir avec une pratique scientifique. En aucun cas nous ne voudrions suggérer que ces trois distinctions partitionnent l'ensemble des usages de langage par les économistes et les agents économiques en catégories nettes et indépendantes. Bien au contraire, c'est lorsque l'on s'intéresse à la manière dont ces catégories interagissent et s'influencent entre elles qu'un certain éclairage peut être apporté sur les débats entre l'économie comportementale et l'économie standard. Par exemple, en économie, l'utilisation de langage formel encourage la combinaison du terme technique « convexe » provenant des mathématiques

[26] Exactement la même conclusion sur l'importance explicite du rôle du langage pour les économistes standards et comportementaux aurait pu être faite en discutant des deux introductions de *The Psychology of Economic Decisions* (Brocas et Carillo 2003b; 2004a).

avec le terme « préférences » provenant du langage quotidien pour former le terme technique « préférences convexes » (au lieu d'utiliser une version plus longue en langage quotidien comme « une préférence pour une quantité moyenne de divers biens sur une quantité extrême d'un seul bien »). L'historien de la psychologie Kurt Danziger (1997) propose de rendre compte des interactions entre les utilisations des langages techniques et quotidiens par les psychologues. Il distingue la « *p*sychologie » qui désigne les usages quotidiens de langage à propos du mental (par exemple, « *je préfère* le chocolat », « j'ai eu immédiatement *peur* ») de la « *P*sychologie » qui désigne les usages techniques de langage par les psychologues dans leurs activités scientifiques (par exemple, « Cette tâche vise à mettre les sujets dans un état de *dissonance cognitive* », « le scanner cérébral montre que les réponses des sujets étaient dues aux *émotions* », la citation des psychologues par Gul et Pesendorfer ci-dessus, etc.). Nous adoptons cette distinction typographique entre la *p*sychologie quotidienne et la *P*sychologie technique, car elle se révèle bien commode pour discuter du problème de l'interdisciplinarité entre économie et *P*sychologie, sous-jacent aux débats entre économie standard et économie comportementale. Enfin, pour mettre un peu plus en avant les interactions entre les langages techniques et quotidiens, nous les considérerons comme deux occurrences d'usages d'un même *langage ordinaire*, une catégorie plus générale qui sera souvent contrastée avec les usages de langage formel.[27]

De plus, la catégorie de langage ordinaire permet de connecter la perspective méthodologique présentée jusqu'à présent avec le dernier problème sous-jacent aux débats entre économie comportementale et économie standard, c'est-à-dire, le problème de la distinction positif/normatif.

[27] Deux remarques s'imposent. Premièrement, parfois sous des noms différents, ces distinctions entre différents types de langage sont examinées par d'autres auteurs en économie (voir surtout Dennis 1982a; b; 1996; 1998a; b; Vilks, 1995; 1998; Weintraub, 1998; 2002; Backhouse, 1998; Mirowski 2002; 2012; Mongin 2001; 2003; 2004; Giocoli Armatte 2004; 2003), comme c'est aussi le cas concernant le statut psychologique des « préférences » et des « choix » en économie et en théorie de la décision (voir par exemple, Pettit, 1991 ; Rosenberg 1992, chap.5-6; Mongin 2011; Guala 2012; Hands 2012a). Dans les deux cas, ces auteurs ne traitent pas des problèmes sous-jacents à l'économie comportementale qui sont au centre de cette thèse. Secondement, en dehors de l'économie, la question de l'interdisciplinarité dans les sciences contemporaines est discutée principalement dans les « *études* d'interdisciplinarité » (par exemple, Klein 2010; Huutoniemi et al. 2010) ou au sujet de « *l'unité ou la désunité* de la science» (par exemple, Oppenheim et Putnam, 1958; Kitcher, 1984). Cat (2013) fait valoir que les deux perspectives sont parfois compatibles, surtout dans le travail de Galison (1999) qui met l'accent sur les interactions entre différents langages à la fois à l'intérieure des, et entre les, disciplines; il convient de noter qu'Heukelom (2009) a utilisé la perspective de ce dernier sur l'économie comportementale.

0.3 Les problèmes du positif/normatif : la position de l'enchevêtrement[28]

Comme le nom du volume mentionné ci-dessus le suggère – *The Foundations of Positive and Normative Economics* - et comme la perspective historique sur l'économie comportementale développée par Floris Heukelom (2014) dans son ouvrage l'atteste, il existe diverses connexions entre la question de l'interdisciplinarité et le problème de la distinction positif/normatif. En dehors de l'économie comportementale, de la théorie de la décision et de certains sous-champs de la méthodologie et de la philosophie de l'économie, le 'problème du positif/normatif' est rarement associé aux modèles de comportements individuels ou à la rationalité individuelle. Il est plutôt associé à des formes collectives de comportements ou de rationalité et à l'évaluation de l'efficacité du marché. À ce niveau, le problème du positif/normatif concerne l'organisation des sous-champs de l'économie en 'économie positive' et 'économie normative'. L'économie positive est censée décrire, expliquer et/ou prévoir des situations économiques, comme par exemple que 'la demande de pommes a augmenté ou augmentera parce que le prix des pommes a diminué ou diminuera', ce qui pourrait être énoncé en théorie du choix du consommateur ou en analyse de la demande. L'économie normative est censée évaluer, recommander et/ou prescrire des situations économiques, par exemple que 'la production de pommes devrait être subventionnée pour accroître le bien-être des consommateurs', ce qui pourrait être énoncé en économie du bien-être, en économie publique et peut être aussi en théorie du choix social. Bien que les modèles de comportements individuels en économie soient considérés à juste titre comme des 'boîtes à outils' utilisées par les économistes pour modéliser des phénomènes sociaux, le problème du positif/normatif se pose également dans ces modèles à travers les réclamations de rationalité qui sont attachées à leurs interprétations. En économie standard, ces modèles ont une dimension positive en ce sens qu'ils décrivent, expliquent et/ou prévoient *le comportement d'individus rationnels*, par exemple que 'si vous préférez les pommes aux oranges et les oranges aux bananes, alors *vous préférez* ou *préférerez* les pommes aux bananes, *parce que* vous êtes rationnel'. Mais *ces mêmes modèles* ont également une dimension normative, dans le sens où ils évaluent, recommandent et/ou prescrivent *la rationalité des comportements individuels*, par exemple que 'si vous préférez

[28]Note de traduction : en anglais, nous avions utilisé une expression qui aurait dû se traduire ici par « thèse de l'enchevêtrement » qui est à nos yeux plus correcte que « position de l'enchevêtrement ». Notre utilisation de 'dissertation' en anglais permettait de ne pas éveiller la confusion entre la présente 'thèse' et l'usage que nous y faisons de la 'thèse' de l'enchevêtrement. Ce n'est plus possible en français, nous avons en conséquence choisi d'utiliser l'expression 'position de l'enchevêtrement'.

les pommes aux oranges et les oranges aux bananes, alors vous *devriez* préférer les pommes aux bananes, ou bien vous êtes irrationnel'.[29]

Est-ce que l'économie positive et l'économie normative peuvent être totalement séparées ? Comment sont-elles articulées ? Ce sont les questions qui sous-tendent le problème du positif/ normatif au niveau de l'articulation entre économie positive et économie normative. Elles ne seront discutées dans cette thèse que brièvement, et uniquement dans le but d'améliorer notre compréhension du problème du positif/ normatif au niveau individuel. À ce niveau, les questions sous-jacentes sont les suivantes. Comment se fait-il que les axiomes de rationalité sur lesquels sont fondés les modèles de comportements individuels aient été considérés *à la fois* comme étant normatifs *et* positifs avant de devenir positivement non plausible à la suite d'expérimentation, tout en demeurant normativement plausible? D'où provient la normativité de ces axiomes et de ces modèles ? Dans cette thèse, ces questions soulevées dans les débats entre économie comportementale et économie standard seront examinées à travers les lunettes de la position de l'enchevêtrement d'Hilary Putnam (un philosophe), Vivian Walsh et Amartya Sen (deux philosophes et économistes). Nous adoptons cette position principalement en raison du rôle central, dans les arguments qui la constituent, de l'articulation entre langage formel et langage ordinaire. Pour comprendre brièvement cette position, il est utile de rappeler que les racines philosophiques du problème positif/normatif en économie sont que « de nombreux économistes prennent pour acquis le principe positiviste que seuls les jugements mathématiques et les jugements de faits - et pas les jugements de valeur - se prêtent à la *discussion rationnelle* » (Mongin, 2006c, p.258, nos italiques; voir aussi Hands 2001, chap.2-3; 2012b). En effet, il semble facilement réalisable *en principe* de parvenir à un accord par la discussion rationnelle sur ce que sont les faits, par exemple que 'le PIB a augmenté de 10 %' ou que 'Nicolas a vendu une bouteille de vin', parce que les données (et le processus de leurs constructions) peuvent être

[29] Voir Hausman et McPherson (2006) ou Hands (2012b) sur les relations entre ces deux niveaux du problème du positif/normatif. Au niveau individuel tout au moins, il convient de noter que la nuance entre 'descriptif" et 'positif' est *parfois* significative en économie. Globalement, le premier signifie 'description de régularités comportementales humaines' et le second 'explication théorique de comportements humains'. Mais ce n'est pas toujours le cas. Bien souvent les deux termes 'positif' et 'descriptif' sont synonymes et prennent *l'une ou l'autre* de ces deux significations. La même ambiguïté peut être soulevée concernant 'normatif' et 'prescriptif', souvent utilisés pour signifier respectivement 'explication théorique de certaines normes de comportements humains' et 'recommandations pour être en conformité avec certaines normes de comportements humains'. Il convient également de noter que le terme 'positif' est souvent associé à la dimension prévisionnelle d'un modèle *par opposition* à la dimension descriptive de ses hypothèses, en suivant Friedman (1953). Dans cette thèse, le sens précis de ces termes sera explicité uniquement lorsque cela apparaitra comme crucial pour le propos et que le contexte ne fournit pas suffisamment d'indices pour déterminer ce sens.

vérifiées par autrui (au moins en principe). Il en va de même pour savoir si les preuves mathématiques de théorèmes et de propositions sont correctes, par exemple que '$2x = 1 \Leftrightarrow x = \frac{1}{2}$', car les étapes d'une preuve peuvent être contrôlées par autrui. Mais cela ne tient pas pour les valeurs, en particulier lorsqu'elles sont de nature éthique, par exemple énoncer que 'porter des jupes courtes est moralement inacceptable' ou que 'les riches devraient être solidaires avec les pauvres', sur lesquelles il semble qu'aucune procédure scientifique ou formelle ne nous permettent de parvenir à un accord. Puis, quoi qu'il en soit, nous avons le sentiment (comme si nous partagions un métajugement de valeur) que nous devrions être libres d'être en désaccord avec les jugements de valeur de n'importe qui, et ce pour une foule de raisons différentes que tout un chacun se plait à entretenir. La contestation de ce principe positiviste et la proposition d'une alternative sont au cœur de la position de l'enchevêtrement.

Adaptée pour l'examen de l'économie comportementale, ce qui est nécessaire parce que ce n'était pas la cible initiale de Sen, Walsh et Putnam, la position de l'enchevêtrement peut être résumée comme suit. Peut importe à quoi réfèrent les scientifiques et les gens dans la vie ordinaire, par exemple, les économistes et les agents économiques, lorsqu'ils parlent de faits particuliers et de valeurs particulières, ces faits et ces valeurs ne peuvent pas être complètement séparés et examinés en isolation les uns des autres. Mais ils peuvent être distingués et leur articulation peut être examinée. De plus, cette articulation ne peut jamais être pleinement comprise sans prendre en compte le rôle des conventions, théoriques pour l'économiste et sociales pour les agents économiques. Les usages du langage ordinaire jouent un rôle crucial dans la manière dont les conventions sociales sont enchevêtrées avec les faits et les valeurs dans notre vie quotidienne. Et symétriquement, les usages du langage formel de la logique, des mathématiques et de la théorie des probabilités jouent un rôle crucial dans la manière dont les conventions théoriques sont enchevêtrées avec les faits et les valeurs dans la science. C'est cette version de la position de l'enchevêtrement qui sera adoptée dans cette thèse pour examiner les contributions théoriques et empiriques constitutives des débats entre économie comportementale et économie standard afin de fournir une meilleure compréhension du problème sous-jacent du positif/normatif dans les modèles de comportements individuels.[30]

[30] La position de l'enchevêtrement rassemble et assemble des arguments préexistants contre l'existence de deux dichotomies. D'une part, deux types d'arguments selon lesquels *la dichotomie fait/convention ne tient qu'en tant que distinction utile* peuvent être dégagés : (1) en provenance de la philosophie des sciences, par les positivistes

Brièvement, la position de l'enchevêtrement est le fruit d'une synthèse de divers domaines de la philosophie (c'est-à-dire, l'éthique, la philosophie du langage, de la logique, des mathématiques, de l'esprit et des sciences) que Putnam (2002; 2004; 2013; 2015) a proposée avant sa mort récente. Les travaux de Putnam (surtout 2002) connectent avec le problème du positif/normatif en économie principalement par le biais des travaux historiques et méthodologiques de Walsh (surtout 1996). Et Sen est pris à la fois par Putnam et par Walsh comme une illustration de la pratique de la théorie économique consciemment effectuée sans les dichotomies, c'est-à-dire faits/conventions ou faits/valeurs (voir surtout Sen 1987; 2002). En retour, Sen a reconnu que les travaux de Putnam et Walsh dans les années 2000 ont en effet caractérisé ses positions historiques et méthodologiques (voir Sen 2005; 2009, note p.357). En ce qui concerne le rôle du langage, trois philosophes jouent un rôle spécifique dans l'arrière-plan de la position de l'enchevêtrement : W.V.O. Quine (dans les travaux de Putnam), Ludwig Wittgenstein (dans les travaux de Putnam et de Sen) et John Austin (dans les travaux de Putnam et de Walsh). Afin d'éviter que la dimension philosophique de cette thèse ne prenne l'ascendant sur ses contributions dirigées vers l'économie, nous ne discuterons pas les travaux de Quine, Wittgenstein et Austin, mais resterons concentrés sur ceux de Putnam, Walsh et Sen (voir aussi les contributions recueillies par Putnam et Walsh 2011).

Pour la même raison, dans cette thèse, la position de l'enchevêtrement sera mobilisée principalement en *illustrant de façon aussi concrète que possible* divers enchevêtrements de faits, de valeurs et de conventions dans les débats entre l'économie comportementale et l'économie standard. Cette stratégie, qui consiste à travailler par illustrations concrètes, semble cohérente avec l'approche philosophique du pragmatisme ou du néo-pragmatisme à laquelle Putnam, Walsh

eux-mêmes, sur les implications des entités non observables dans la physique quantique, avec aussi le travail de Quine depuis « Two Dogmas of Empiricism » (1951); et (2) en provenance de la philosophie des mathématiques et de la logique à propos de l'indispensabilité des mathématiques et de la logique en science, et aussi à propos de l'accord consensuel sur leur 'objectivité' malgré leur absence de référence à des objets observables dans le sens selon lequel la science est censée référer aux objets observables selon le positivisme (voir surtout Putnam 1971; 2004). D'autre part, deux types d'arguments selon lesquels *la dichotomie fait/valeur ne tient qu'en tant que distinction utile* peuvent être dégagés: (1) en provenance de l'école dite du pragmatisme en philosophie qui fait valoir qu'en dehors des jugements de valeur éthiques, la pratique de la science nécessite beaucoup de *jugements de valeur épistémiques* via des notions comme la 'simplicité', la 'cohérence', la 'pertinence', la 'conservation des théories passées' dans la construction et dans la sélection des théories (voir surtout Putnam 2002, chap. 2); et (2) en provenance des débats d'après 1960 en méta-éthique sur la distinction entre les prédicats fins permettant de porter des jugements de valeur purement évaluatifs (ou 'normatifs'), comme par exemple 'bon', 'juste', 'il est moralement bon de ne jamais mentir', et les *prédicats épais* permettant de porter des jugements de valeur qui sont à la fois, et de façon inséparable, évaluatifs et descriptifs (ou 'normatifs' et 'positifs'), comme par exemple 'cruel', 'rude', 'rationnel', 'préfère', 'il est rationnel de vendre une bouteille de vin lorsque l'on ne boit pas de vin'.

et Sen sont souvent identifiés et pour laquelle ils ont explicitement manifesté de la sympathie. Notamment, cette stratégie semble correspondre aux premiers traits utilisés par Wade Hands (2001, pp.216-217) pour caractériser le pragmatisme ou le néo-pragmatisme : (1) le refus d'une méthodologie fondationnaliste selon laquelle les critères pour l'évaluation de contributions scientifiques sont définis *a priori* et de l'extérieur de la pratique scientifique, et (2) le refus d'une dichotomie tranchante entre contributions théoriques et contributions empiriques dans une discipline donnée. De plus, cette stratégie semble également adaptée aux contributions proposées dans la thèse, comme nous l'expliquons dans la section suivante.[31]

0.4 Contributions proposées et articulation des chapitres

Pour mieux situer la perspective méthodologique, sur le rôle du langage dans la rationalité économique, présentée jusqu'à présent, nous pouvons la comparer à deux autres positions existantes sur cette question en méthodologie et en philosophie de l'économie. D'une part, il y a la position de Deirdre McCloskey (1998 [1985]), qui a analysé la rhétorique sous-jacente aux utilisations du langage par les économistes. Brièvement, McCloskey soutient que (1) les normes de conversation dans le langage quotidien sont nécessaires et suffisantes pour régler (2) les problèmes méthodologiques que rencontrent les économistes. Par conséquent (3) les prescriptions épistémologiques plus ou moins inspirées de la philosophie des sciences que proclament parfois les économistes, les méthodologues ou les philosophes, produisent des interférences parasites entre (1) et (2). D'autre part, il y a la position de Don Ross (2005; 2014), qui peut être considérée comme étant à l'exact opposée de celle de McCloskey; et à la différence de cette dernière, il traite de l'économie comportementale. Brièvement, Ross soutient que (1) le langage quotidien produit des interférences parasites dans la relation entre (2) les questions méthodologiques autour de l'économie comportementale et (3) les solutions philosophiques qu'il y propose à partir de la philosophie de l'esprit et des sciences (voir Ross, Ladyman et Spurrett et 2007 pour une version plus générale de sa position). La perspective méthodologique sur le rôle du langage dans la rationalité économique développée jusqu'ici, en particulier telle que nous l'avons ancrée dans la

[31] Il convient de noter que le trait (1) refuse également le relativisme radical. Les différences entre pragmatisme et néo-pragmatisme ne sont pas significatives pour l"objectif de cette thèse, mais notons tout de même que ces positions ne doivent pas être confondues avec la pragmatique du langage discuté au début de cette introduction générale, malgré d'existantes connections non-triviales (voir Hands 2001, chap.6).

position de l'enchevêtrement, se trouve entre ces deux extrêmes. En bref, cette thèse soutient que c'est parce que (1) le langage quotidien (en particulier la *p*sychologie) est une dimension indispensable *à la fois* des choix que font les agents économiques et des théories que proposent les économistes, que lui accorder une certaine attention peut contribuer *à la fois* à (2) et à (3). Cela peut contribuer à (3) certains problèmes philosophiques posés par les méthodologues ou philosophes s'intéressant à la théorie économique ou posés par les économistes eux-mêmes dans leurs manifestes méthodologiques. Et cela peut également contribuer à (2) certains problèmes méthodologiques, théoriques et empiriques rencontrés par les économistes s'intéressant aux comportements individuels des agents économiques.

Pour résumer, la contribution proposée dans cette thèse est de démontrer comment une perspective méthodologique sur le double rôle du langage dans la rationalité économique peut clarifier trois principales questions (et leurs connexions) qui sous-tendent les débats entre économie comportementale et économie standard : le problème de l'unification théorique vis-à-vis des trois dimensions de la rationalité économique, la question de l'interdisciplinarité entre économie et *P*sychologie, et le problème du positif/normatif dans les modèles de comportements individuels. De plus, cette thèse vise à aller au-delà de la simple clarification concernant le problème du positif/normatif et du rôle du langage dans les comportements des agents économiques. Notre intention est de fournir une critique constructive des contributions de l'économie comportementale, ainsi que celles de l'économie standard sur ces deux points. Suivant la position de l'enchevêtrement, il sera soutenu que l'économie tant standard que comportementale propose une articulation insatisfaisante des dimensions positive et normative dans les modèles de comportements individuels; et que la reconnaissance de l'enchevêtrement de faits, de valeurs et de conventions peut se révéler être théoriquement et empiriquement fructueuse. Il sera également soutenu que prêter attention au rôle du langage dans les comportements des agents économiques montre *parfois* qu'un comportement apparemment irrationnel peut en fait être défendu comme rationnel; c'est pourquoi nous soutiendrons que l'axiome implicite - connu sous le nom d'invariance à la description - dans les modèles standards de comportements individuels empêchant l'influence du langage doit être affaibli (mais pas complètement supprimé), contrairement aux positions de la plupart des économistes standards et comportementaux.

Cette thèse est structurée en cinq chapitres, dont l'articulation est expliquée dans le résumé

des développements qui suit.

Résumé étoffé du développement

Le premier chapitre développe 'en action' la perspective méthodologique esquissée dans l'introduction générale, c'est-à-dire que les développements sont construits à partir des problèmes concrets qui sous-tendent les débats entre économie comportementale et économie standard. L'ancrage de cette perspective dans la position de l'enchevêtrement est également développé 'en action', en discutant les critiques adressées par Sen aux mêmes modèles standards de comportements individuels qui sont critiqués par l'économie comportementale. Le chapitre est structuré autour d'une étude comparative des travaux de Sen (représentant la position de l'enchevêtrement) et Thaler (représentant la position de l'économie comportementale). Nous nous concentrons sur la théorie du choix du consommateur et la théorie du choix rationnel *en situation de certitude*, sur lesquelles portent les contributions principales des deux auteurs. Cela revient à comparer deux critiques différentes des mêmes modèles de comportements économiques sous certitude. Alors que les deux critiques prennent comme point de départ une observation commune, qui est la suivante : les dimensions positive et normative des modèles standards de comportements individuels doivent être articulées différemment. Les deux critiques aboutissent néanmoins à des conclusions radicalement différentes. Au travers d'une relation interdisciplinaire avec l'éthique, la philosophie morale et la philosophie politique, Sen propose de développer la dimension normative à partir d'un examen minutieux de faits empiriques; c'est-à-dire que la rationalité ne doit pas être définie *a priori* et une fois pour toutes. Au travers d'une relation interdisciplinaire avec la *P*sychologie, Thaler propose de séparer la dimension normative de la dimension positive; c'est-à-dire que les modèles standards fournissent une notion adéquate de la rationalité qui peut être utilisée comme référence pour évaluer les comportements empiriques.

Les deux points principaux de ce chapitre concernant le développement général de la thèse sont les suivants. D'une part, si l'on veut prendre au sérieux le rôle du langage dans les choix des agents économiques, alors ce que nous appelons *la structure communicative des choix* doit être rendu explicite. Nous proposons de faire cela de deux manières complémentaires. La première consiste à distinguer deux entités : le *modeleur de décision* qui pose un problème de

décision à un *preneur de décision* qui effectue un choix. Le modeleur de décision peut, parmi d'autres possibilités, être un économiste, en particulier dans les expériences de laboratoire où les questions liées aux interactions entre le langage de l'économiste et celui de l'agent économique ne sont pas triviales. La seconde consiste à introduire la théorie des actes de langage (autrement appelée théorie des actes de parole), notamment dans la version qu'en propose le philosophe John Searle, afin d'étoffer la structure communicative d'une façon qui est pertinente pour la théorie du choix rationnel. D'autre part, alors que nous nous concentrions sur l'analyse économique sous certitude, les trois dimensions des comportements individuels (l'incertitude, le temps et autrui) n'ont cessé de surgir lorsque des questions liées aux causes ou facteurs déterminants les préférences individuelles se posent (c'est-à-dire, lorsque les préférences ne sont plus prises comme étant 'données').

Dans le deuxième chapitre les défis, dans les trois dimensions des comportements individuels, posés par l'économie comportementale à l'économie standard sont examinés. Nous structurons notre discussion autour d'une rupture historique entre deux périodes. La première période débute à la fin des années 1970 et s'étend jusqu'au milieu des années 2000. C'est la période des défis classiques posés par l'économie comportementale *dans* les trois dimensions, indépendamment les unes des autres. Nous proposons, pour les trois dimensions, un examen comparatif (1) des régularités empiriques constituant ces défis, (2) des modèles standards mis en défaut, et (3) des alternatives de l'économie comportementale. Concernant le problème de l'interdisciplinarité, l'examen de cette période révèle des inspirations très claires provenant de la *P*sychologie : les travaux de Kahneman et Tversky dans la dimension de l'incertitude, et les travaux de George Ainslie dans la dimension du temps. Cela marque un contraste fort avec la dimension du rapport aux autres, pour laquelle aucune forme d'interdisciplinarité claire, avec par exemple la *P*sychologie sociale, n'existe. Nous insistons, à propos du problème positif/normatif, sur la manière dont les trois dimensions sont liées par la doctrine du conséquentialisme. Doctrine à partir de laquelle les économistes et les psychologues dérivent leurs jugements de valeur à propos de la rationalité ou l'irrationalité des comportements. La seconde période est contemporaine et commence au milieu des années 2000. Dans celle-ci, nous observons l'émergence d'un ensemble de défis ayant trait aux interactions *entre* les dimensions, et posant des problèmes *à la fois* aux modèles de l'économie standard et aux alternatives de l'économie comportementale.

Trois points de ce chapitre peuvent être dégagés vis-à-vis du développement général. Le premier est un développement de la structure communicative des choix : que ce soit pour les interactions dans les dimensions ou entre les dimensions, une condition de possibilité des défis qui sont discutés dans le chapitre est le marquage linguistique de distinctions relatives aux trois dimensions (par exemple, 'maintenant' et 'plus tard', 'certainement' et 'probablement', 'toi' et 'moi') que le modeleur de décision (par exemple, l'économiste conduisant une expérience de laboratoire) doit faire, et que le preneur de décision (par exemple, l'agent économique participant à l'expérience) doit comprendre. Le deuxième point concerne la normativité du conséquentialisme en théorie de la décision. Historiquement, il y a eu une lente transformation par laquelle la primauté initiale de la dimension de l'incertitude (sur les deux autres dimensions) dans les jugements de valeur de rationalité ou d'irrationalité des comportements individuels, s'est faite concurrencer par une primauté de la dimension du temps. Concrètement, cela s'observe dans le passage d'arguments qualifiés d'intuitifs, pour l'axiome d'indépendance de l'utilité espérée, à des arguments fondés sur la notion de cohérence dynamique. Enfin, le troisième point concerne l'unification théorique. Etant donné que les régularités comportementales entre les dimensions, c'est-à-dire les défis de la seconde période, posent des problèmes à la fois pour l'économie standard et pour l'économie comportementale, l'unification théorique recherchée par les économistes standards comme comportementaux ne peut ignorer les interactions existantes entre les dimensions, impliquant de ne plus considérer les trois dimensions comme étant déconnectées les unes des autres. En bref, les interactions entre les trois dimensions représentent un domaine potentiellement fructueux pour une éventuelle convergence entre économie comportementale et économie standard, au moins concernant le problème de l'unification théorique.

Dans le troisième chapitre, est examiné un ensemble de modèles cherchant explicitement l'unification théorique entre économie comportementale et standard, que nous appelons les « modèles duaux ». Cet examen est réalisé dans le but de savoir si les modèles capturent ou non les interactions entre les trois dimensions. Le rôle du langage dans la rationalité économique, telle qu'elle est construite par les économistes dans leurs modèles, est étudié en faisant abstraction de l'utilisation du langage par les agents économiques dans leurs comportements. Plus précisément, nous étudions les relations entre, d'une part, les usages de langage formel et technique ayant trait à l'économie, et, d'autre part, un nouveau langage technique emprunté à la *P*sychologie

du contrôle de soi (*self-control*). Parce que les problèmes de contrôle de soi sont principalement liés à la relation du preneur de décision avec le temps, nous suggérons que l'évolution historique d'une primauté de la dimension de l'incertitude à une primauté de la dimension du temps, observée dans le chapitre précédent, trouve son illustration paradigmatique dans les modèles duaux. Nous montrons également qu'un seul de ces modèles, à savoir, celui de Fudenberg et Levine's, capture une partie des interactions entre les trois dimensions, fournissant ainsi une avenue prometteuse d'unification théorique et de réconciliation entre économie standard et économie comportementale.

Deux points concernant le développement général de la thèse peuvent être dégagés. Premièrement, Fudenberg et Levine prennent une version affaiblie de la position standard concernant le problème positif/normatif dans leur modèle dual, bien que ce dernier ne soit pas entièrement standard. C'est-à-dire qu'ils considèrent que leur modèle peut, *dans certains cas où il contredit le modèle standard*, s'interpréter *à la fois* de manière positive et normative. Nous soutenons qu'une vision empiriste de la rationalité sous-tend cette position, selon laquelle la rationalité individuelle ne doit pas tenir lieu que d'idéal, c'est aussi un phénomène empiriquement observable. Une implication est que la notion de cohérence, utilisée pour qualifier les modèles standards de normatifs, se trouve ici modifiée en considérant que les limites psychologiques responsables des problèmes de contrôle de soi sont une contrainte inévitable sous laquelle les agents économiques optimisent. Secondement, un ensemble de régularités comportementales connu sous le nom de d' « effets de cadrages » est explicitement pointé par Fudenberg et Levine (avec d'autres auteurs et d'autres modèles duaux) comme la principale limite à l'unification théorique.

Par conséquent, dans le chapitre 4 nous dirigeons notre attention vers ce que les économistes standards comme comportementaux énoncent à propos du « cadrage », et vers ce que nous pouvons tirer d'une trentaine d'années d'études sur ce phénomène par les psychologues pour en comprendre la structure empirique et les implications théoriques. Nous avons tout d'abord observé que ce terme est utilisé pour désigner un ensemble de plus en plus hétérogène de phénomènes comportementaux. Le seul point commun de ces phénomènes semble être que différentes présentations du 'même' problème de décision peuvent révéler des renversements de préférences ; le terme 'même' doit être compris comme 'le même du point de vue des modèles standards de comportements individuels'. La seule attention que nous portons au rôle du langage dans la

rationalité économique, telle qu'elle est construite par les économistes, consiste à souligner le rôle crucial joué par les conventions théoriques des modèles standards dans l'établissement de relations d'équivalence entre problèmes de décision. Nous avons décidé d'étudier les effets de cadrage en nous concentrant sur le sous-ensemble de ceux qui violent l'axiome implicite des modèles standards, « l'invariance à la description », qui rend ces modèles aveugles aux usages de langage par les agents économiques ; ainsi le rôle du langage dans la rationalité économique, telle qu'elle s'expose dans les comportements des agents économiques, est étudié en faisant relativement abstraction de l'utilisation du langage par les économistes. En dépit de certaines études sur les violations de l'invariance à la description par des économistes comportementaux, cet axiome demeure relativement peu étudié en économie compte tenu de l'important corpus de recherches sur le sujet qui existe en *P*sychologie. Le chapitre propose d'organiser ce que nous considérons comme étant les principales caractéristiques, du point de vue d'un économiste, de trois décennies de recherche par les psychologues sur les violations de l'invariance à la description.

Le principal point concernant le développement général de la thèse est le suivant. La vision générale que les économistes ont des effets de cadrage, fondée sur les travaux pionniers de Kahneman et Tversky dans les années 1980 et ignorant trois décennies de recherche par d'autres psychologues, est insatisfaisante. Cette vision consiste à prendre l'axiome d'invariance à la description comme une condition ultime de rationalité, dont les violations seraient toujours irrationnelles. Or, en examinant les travaux des autres psychologues, deux conditions sous lesquelles les violations de l'invariance à la description sont potentiellement rationnelles peuvent être dégagées. Premièrement, les préférences peuvent dépendre *directement et consciemment* des descriptions, de sorte que l'on peut considérer qu'il y a maximisation d'une fonction d'utilité lors du choix d'une description d'un objet, mais pas lors du choix d'une autre description du même objet. Secondement, la structure communicative des choix fait qu'il est possible que le choix d'une description, et non d'une autre d'un objet, par le modeleur de décision soit une source d'information plus ou moins tacite à propos du problème de décision, et que le preneur de décision infère rationnellement cette information. En bref, il existe des raisons de ne pas considérer l'axiome d'invariance à la description comme un gage ultime de la rationalité, ce qui peut justifier la volonté d'affaiblir cet axiome pour capturer, non seulement les phénomènes de cadrage, mais également la possible rationalité qui leur est sous-jacente.

Le dernier chapitre est coécrit avec un praticien de la théorie de la décision axiomatique, Dino Borie. Nous proposons une axiomatique formelle dans laquelle l'axiome d'invariance à la description peut être rendu explicite d'une manière assez fine pour ensuite pouvoir explorer différents affaiblissements de cet axiome qui rendent compte des effets de cadrage. La contribution théorique du chapitre est adressée à l'économie en se fondant sur les résultats des psychologues étudiés au chapitre précédent. Nous appliquons cette axiomatique à différents effets de cadrage pour montrer que les modèles existants (mais peu nombreux) pour le cadrage en économie ne traitent pas des violations de l'invariance à la description, ainsi que du positif/normatif sous-jacents aux effets de cadrage, de manière satisfaisante.

Les principaux points concernant le développement général de la thèse sont les suivants. Tout d'abord, notre interprétation de la structure mathématique de ce travail formel est systématiquement proposée en termes de deux perspectives distinctes, mais possiblement connectées, sur un même problème de décision : celle du modeleur de décision et celle du preneur de décision. Cette interprétation vise à mettre en avant la structure communicative des choix. Ensuite, nous soutenons que, contrairement à la vision générale des économistes sur le cadrage, l'axiome d'invariance à la description n'est pas un axiome pour lequel un choix binaire s'impose entre travailler avec ou travailler sans. Nous montrons comment il est possible de garder une version affaiblie de cet axiome sans s'en débarrasser totalement, ce que nous faisons de deux manières, qui elles-mêmes impliquent deux degrés différents de dépendance à la description. Cette stratégie nous permet d'apporter quelques clarifications au problème du positif/normatif sous-jacent les violations de l'invariance à la description, en qualifiant les conditions axiomatiques sous lesquelles des interprétations en termes de rationalité ou d'irrationalité peuvent s'appliquer à un effet de cadrage donné. Enfin, nos affaiblissements de l'axiome d'invariance à la description sont conduits de sorte que notre axiomatique puisse être intégrée dans les modèles comportementaux, mais aussi standards. Pour ces derniers, nous démontrons comment l'équivalence standard entre fonction de choix, relation de préférence et fonction d'utilité tient dans notre axiomatique. La simplicité de notre cadre et la démonstration de ces équivalences sont conçues comme des modifications mineures dans le langage formel et technique des économistes, qui permettent de prendre en compte des utilisations du langage quotidien par les agents économiques qui ne sont pas triviaux d'un point de vue économique.

Conclusion générale

La contribution générale que nous avons tenté d'apporter dans cette thèse consiste à montrer comment une perspective méthodologique sur le double rôle du langage dans la rationalité économique peut clarifier trois principaux problèmes (et leurs connexions) qui sous-tendent les débats entre économie comportementale et économie standard. Il s'agissait du problème de l'unification théorique dans les trois dimensions (incertitude, temps, autrui) de la rationalité économique, du problème de l'interdisciplinarité entre économie et *P*sychologie, et du problème du positif/normatif dans les modèles de comportements individuels. Nous avons discuté plusieurs marqueurs de la frontière entre économie standard et économie comportementale. Des marqueurs théoriques, qui sont constitués d'une série de notions à la fois psychologique et Psychologique telles que l'aversion à la perte sous certitude comme sous risque, l'impulsivité dans le temps, et diverses motivations non autocentrées envers autrui. Des marqueurs empiriques, qui sont essentiellement constitués des effets de cadrage. Par ailleurs, il n'y a pas véritablement de marqueurs normatifs, étant donné que la doctrine du conséquentialisme est largement acceptée par les deux côtés pour dériver des jugements de valeur de rationalité et d'irrationalité à propos de ce qui se trouve du côté de l'économie comportementale, c'est-à-dire à propos des marqueurs théoriques et empiriques. Une des implications de cela est, par exemple, que les débats entre économie comportementale et économie standard sont aveugles à la notion d'engagement défendue par Sen. Nous avons également défendu qu'une faiblesse de cette position conséquentialiste soit que la notion théorique de 'conséquence', utilisée par la plupart des économistes, ne semble pas être assez précise pour être l'objet d'une science empirique.

Cette thèse a tenté d'aller au-delà d'une simple clarification concernant le problème du positif/normatif et le rôle du langage dans les comportements des agents économiques. L'intention était de proposer, sur ces deux points, une critique constructive des contributions de l'économie comportementale comme standard. En suivant la position de l'enchevêtrement (développé par Putnam, Sen et Walsh), nous avons soutenu qu'à la fois l'économie standard et l'économie comportementale proposent une articulation insatisfaisante entre les dimensions positive et normative des modèles de comportements individuels; et que la reconnaissance de l'enchevêtrement de faits, de valeurs et de conventions peut se révéler être théoriquement et empiriquement

fructueuse. En outre, il a été soutenu que prêter attention au rôle du langage dans les comportements des agents économiques montre parfois qu'un comportement apparemment irrationnel peut en fait être défendu comme rationnel. En conséquence, nous avons défendu que l'axiome implicite - connu sous le nom d'invariance à la description - dans les modèles standards de comportements individuels empêchant l'influence du langage doive être affaibli (mais pas complètement supprimé), contrairement aux positions de la plupart des économistes standards et comportementaux. La contribution formelle sur ce point était en partie destinée à montrer comment des réflexions en provenance de la méthodologie et de la philosophie de l'économie peuvent avoir une influence bénéfique sur la théorie économique.

Concernant les travaux à venir, certains projets en lien avec cette thèse sont déjà en cours. Les arguments esquissés dans le chapitre 2, sur la possibilité de réinterprétations critiques d'études empiriques existantes ayant utilisé une seule dimension dans l'interprétation originale de leurs données, sont développés dans une série d'études de cas avec Judith Favereau et Cléo Chassonnery-Zaïgouche. D'autres arguments, également esquissés dans le chapitre 2, sur les relations entre narrativité et identité vis-à-vis de la cohérence dynamique, sont l'objet d'un travail entrepris avec Tom Juille. Certains des arguments esquissés dans les chapitres 1 et 2, sur la traduction expérimentale de la rationalité, sont développés dans une perspective épistémologique et historique dans un document de travail ; le point de vue épistémologique tente à la fois d'ancrer la traduction expérimentale de la rationalité dans la théorie de l'équilibre réfléchi tout en fournissant simultanément une possible opérationnalisation empirique de cette dernière; la perspective historique discute l'influence (due à l'émergence de l'économie comportementale) de l'économie sur la *P*sychologie dans les débats sur la rationalité en sciences cognitives. Lors de l'élaboration du document de travail qui sous-tend le chapitre 5, nous avons, avec Dino Borie, fait quelques expériences sur nos étudiants afin de tester certains de nos axiomes; nous sommes actuellement en train de discuter de la possibilité de faire ces expériences en laboratoire avec les expérimentalistes de l'Université de Nice Sophia-Antipolis ; notamment Ismaël Rafaï avec qui nous avons également discuté de la possibilité de construire un modèle d'effets de cadrage à partir de notre cadre axiomatique.

Nous proposons de conclure cette thèse avec Fritz Machlup. Dans un papier intitulé « Que se passerait-il si la matière pouvait parler », Machlup introduit le problème méthodologique de la

distinction entre les sciences sociales et les sciences naturelles avec une histoire, dont le passage suivant est tiré :

> « Alors qu'il parlait de la marche aléatoire des molécules et des collisions moléculaires sous différentes pressions, quelqu'un a crié « Arrêtez ces absurdités ! ». Quand il a regardé autour pour voir quel étudiant avait fait cette remarque impertinente, cette voix continuait de se faire entendre. Elle provenait de toute évidence de la chambre de protection avec le miroir suspendu, dont les déplacements étaient suivis par la fluctuation d'un faisceau de lumière réfléchie. Voici ce qu'il entendit : « Il est temps que vous cessiez d'induire vos étudiants en erreur. Ce que vous enseignez à propos de nous les molécules n'est tout simplement pas vrai. Ce n'est pas une marche aléatoire et nous ne nous poussons pas les unes sur les autres un peu partout et dans n'importe quel sens. Nous savons où nous allons et pourquoi. Si vous voulez écouter, nous serions ravies de vous le raconter. » Il n'a pas attendu un instant de plus, s'est précipité pour rapporter tout cela, et ramena le professeur R. pour assister à l'événement afin d'entendre ce que les molécules étaient sur le point de raconter. « Oh, » a déclaré le professeur R., « Vous voulez dire qu'elles vont nous raconter ce qu'elles *pensent* qu'elles font. Par tous les moyens, laissez-les faire. » » (Machlup 1978, pp.310-311)

L'objectif de Machlup (1978 [1969]) est de défendre qu'une des spécificités des sciences sociales est que leurs « données et problèmes » (ibid, p.319) sont imbibés d'usages de langage, non seulement venant des scientifiques (ce qui n'est pas une spécificité des sciences sociales), mais aussi venant de ses objets d'étude. Plus précisément, il soutient que les sciences sociales sont spécifiques en raison de la possibilité de « communications contradictoires » (ibid.) entre les scientifiques et leurs sujets. Alors qu'un grès ne peut pas dire à un géologue « Non, nous ne sommes *pas* de cette famille que vous appelez « roches sédimentaires », nous sommes très différents de ces individus », un preneur de décision peut dire à un théoricien de la décision « Non, je ne suis pas comme vous le dites « irrationnel » ou « quasi-rationnel », j'ai de nombreuses raisons de maintenir mes préférences comme elles sont, et je ne les « renverse » pas ». Comme nous l'avons vu, c'est à peu près ce qui s'est passé autour du paradoxe d'Allais durant une courte période de l'histoire à la fin des années 1970. Des scientifiques de la décision (MacCrimmon, Larson, Slovic et Tversky entre autres) *ont écouté* leurs sujets au lieu de simplement chercher à observer leurs comportements; et cela a contribué à des développements théoriques non triviaux. Comme nous l'avons également vu, cela n'est pas dénué de dangers méthodologiques, notamment parce que de nombreuses caractéristiques subtiles de la *conversation* ne peuvent être contrôlées, de la même manière que la *communication* minimaliste dans des expériences de laboratoire peut

l'être.

De toute façon, l'utilisation du langage dans la rationalité économique, qu'elle soit construite par les économistes ou exposée par les comportements des agents économiques, y comprit dans la structure communicative des choix qui peut réunir ces deux entités, elle n'est pas restreinte au problème de la communication contradictoire. Il est plus facile d'énoncer les problèmes qui nous ont préoccupés en inversant la question de Machlup : que se passerait-il si les agents économiques ne pouvaient plus parler ? Si tel était le cas, alors, comme nous l'avons soutenu au chapitre 2, aucun des défis classiques ou récents posés par l'économie comportementale dans les trois dimensions de la rationalité économique n'aurait pu être posé en premier lieu; comme nous l'avons soutenu au chapitre 4, les effets de cadrage n'existeraient pas; et si nous gardons à l'esprit que les économistes sont également des agents économiques, alors, comme nous l'avons soutenu au chapitre 1 et au chapitre 3, la construction de modèles de comportements individuels serait gravement compromise, voire totalement impossible. Un avertissement que nous avons reçu à propos de notre focalisation sur le rôle du langage était à peu près le suivant. 'N'y vas pas, dès que tu commences à t'intéresser au langage, tu vois du langage partout !' Nous sommes non seulement en accord avec le contenu de cet avertissement, mais également avec l'existence des dangers méthodologiques réels qui le sous-tendent. Néanmoins, si le langage n'est pas, *stricto sensu*, *omniprésent*, il est tout de même prévalant d'une manière non triviale pour l'économie. Selon nous, il y a tout autant de dangers méthodologiques à nier la prévalence d'un phénomène prévalant.

Remerciements and Acknowledgments

Comme d'habitude, quand j'écris un texte et que personne ne passe derrière, c'est long, trop long. Et puis c'est souvent un peu lourd aussi, oui bon... souvent, ça ne l'est pas qu'un peu. Mais là c'est pour parler de vous alors ça devrait être supportable ; et puis vous pouvez toujours faire Ctrl+F si ça ne l'est pas. Il y a tellement de personnes que je souhaite remercier que mes premiers remerciements vont à ceux que je risque d'oublier... S'il vous plait, mettez cela sur le compte de la fatigue de fin de thèse, et considérez ce premier paragraphe comme l'émission d'un bon valable à vie et pour le nectar de votre choix lors de notre prochaine rencontre.

Il y a quatre personnes que je souhaite remercier par ordre chronologique des rôles déterminants qu'ils ont joués dans mon orientation vers cette aventure doctorale. La première c'est François Dumarest. Merci d'avoir un jour mis dans ma tête cette idée, à l'époque assez étrange, de non seulement continuer mes études sur Paris, mais aussi de regarder ces choses abstraites que sont la recherche et le doctorat comme 'des trucs cool'. La deuxième c'est Annie Cot. Tout d'abord, merci d'avoir accepté, en plus de bien vouloir faire partie du jury, cette tâche pas forcément agréable de rapporter la thèse ; mais merci surtout pour ce mélange de détachement et de rigueur que tu parviens à transmettre d'une manière – je vais le dire avec ces mots que toi seule sais utiliser – 'chic et délicieuse'. Juste après Annie, ou plutôt juste en face d'elle, se trouve Jérôme Lallement. Merci encore pour cette transmission d'un mélange de rigueur et de détachement, avec également (cela s'applique aussi à Annie) l'inspiration et l'envie que procure vos cours, et – là ce n'est que pour vous – la lecture de la votre de thèse. Enfin il y a mon directeur de thèse, Richard Arena. Si ces cinq années de thèse ont été une énorme partie de plaisir intellectuel, c'est en grande partie grâce à toi. Merci pour ce style – que je m'aventurerai à qualifier de Wittgensteinien – dans la manière de dire et la manière de faire – si cette distinction a un sens ; c'était, dans l'ensemble, vraiment très agréable.

Je voudrais ensuite remercier ceux qui vont officiellement permettre (enfin !) que tout cela se termine : le jury. Merci Agnès Festré pour toutes les conversations que nous avons pu avoir depuis le début de la thèse ; tes encouragements ont clairement contribué à entretenir ma motivation (notamment *via* un souvenir qui me revient souvent en mémoire: les premiers retours que tu m'avais très sympathiquement donnés sur mon premier papier sans que je ne demande rien). Merci Christian Hudelot pour cette érudition humble que tu transmets avec générosité dans le groupe de lecture ; elle n'a pas laissé la trajectoire de cette thèse inchangée. Je remercie sincèrement Guillaume Hollard de bien avoir voulu s'infliger la lecture du manuscrit ; le peu d'échanges que nous avons eu par le passé a durablement marqué la manière dont je conçois l'articulation entre la réflexion méthodologique et la pratique de l'économie. I sincerely thank Matthias Klaes for having accepted to be part of the jury; I am very excited by the opportunity to discuss methodology and behavioral economics with such an important contributor to these topics. And for the same reasons, I warmly thank Wade Hands for having accepted to be part of all this, especially for the referring duties... Wade, the next one is definitely on me!

Keeping it in English for the sake of harmony, there are a few people whom I would like to thank in that language. First of all, John Davis, thank you so much for the humanism you display in your interpersonal relations, for the encouragements, advices and constructive criticisms over the years. Then, Niels Geiger, thank you for all those passionate conversations on behavioral economics and for your acute reading of chapter 5. Nuño Martins, thank you for those passionate conversations on the entanglement thesis and economic rationality (and for your kind words at the very beginning of the dissertation, they marked my mind). Kyu Sang Lee, thank you for those passionate conversations on experimental economics. Robert Sugden, thank you for this conversation on experimental and behavioral economics' "behind the scenes". Scott Scheal, thank you for these nice conversations on economics and philosophy, and also for having slightly reassured me on the English of the dissertation. Tom Cunningham, thank you so much for your generosity, all these comments and criticisms you gave me by email (without even knowing me) were very helpful; and thank you, Anna Dreber, for your generosity and for encouraging me to get in touch with Tom. I would also like to make two general thank. One for some of the actors of this dissertation who took the time to answer my annoying emails with some encouraging words and/or some useful comments and criticisms on some chapters: Colin Camerer, Andrew Caplin, Andrew Schotter, David Levine, Alberto Bisin, Juan Carrillo, Peter Hammond, Slomi Sher and especially Peter Wakker. And another one for the philosophers and economists of the 2014 INEM/CHESS Summer School for some vey insightful interactions: Julian Reiss, Conrad Heilmann, Till Grüne-Yanoff, Nathalie Gold, Alan Kirman, Uri Ansenberg, Goreti Faria, and especially Raj Patel. And last but not least, thank you Brandon Unti for all the passionate conversations about life and (a bit of) economics and philosophy during my coolest road trip experience ever!

Il y a un ensemble de chercheurs qui, je le pense sincèrement, ont largement influencé le contenu de cette thèse. Il y a évidemment les coauteurs. Nicolas Vallois, merci de m'avoir proposé de faire un papier au tout début de la thèse, cela a largement influencé ma manière de travailler pour la suite. Judith Favereau et Cléo Chassonnery-Zaïgouche, merci de m'avoir proposé de faire un papier que l'on va maintenant pouvoir terminer à la bien ; mais merci surtout pour vos styles respectifs qui impactent le mien d'une manière délicieusement conflictuelle. Dino Borie, merci de m'avoir proposé de travailler avec toi, cela a eu une influence majeure et évidente dans les deux derniers chapitres de la thèse, mais aussi sur mon mode de raisonnement de manière bien plus générale. Tom Juille, merci d'avoir accepté de faire un papier avec moi, et pour toutes ces conversations quotidiennes dont la philosophie ordinaire ne laisse personne indifférent (et puis merci pour le nombre conséquent de relectures au cours des trois dernières années). A côté des coauteurs (mais peut-être pas si loin), et pour rester sur l'importance des conversations quotidiennes, merci Lauren Larrouy (pour pleins de choses, et particulièrement pour l'intuition derrière l'articulation des chapitres 3 et 4). Pour les mêmes raisons, merci Nicolas Brisset, particulièrement de m'avoir encouragé à ne pas bâcler l'affaire.

Je dois également une série de profonds remerciements à ceux que je n'ai pas encore cités de la joyeuse bande des séminaires Albert O. Hirschman auxquels j'ai assisté : Agnès Gramain, Jean-Sébastien Lenfant, Niels Boissonnet, Pierrick Dechaux, Aurélien Goutsmedt, Erich Pinzon Fuchs, Francesco Sergi, Isselmou Oud Boye, Juan Melo, Quentin Couix, et puis mes deux compagnons de route, Maxime Desmarais-Tremblay et Matthieu Renault. Pour les mêmes raisons, je voudrais remercier deux des anciens de ces séminaires qui ont notablement influencé des parties de cette thèse : merci Samuel Ferey et merci José Edwards. J'adresse aussi un remerciement collectif à une autre joyeuse bande qui a influencé cette thèse sur d'autres aspects, celle des expérimentalistes des « Friday Meetings » : Nobi Hanaki, Eric Guerci, Michela Chessa, Mira Toumi, Imen Bouhlel, Sébastien Duchêne ; avec deux mercis un peu plus particuliers du fait de leurs influences respectives un peu plus spécifiques, merci Pierre Garrouste et merci

Ismaël Rafaï, pour cette attitude dont vous avez le secret, dans la recherche comme dans la vie.

La liste des individus que j'ai découvert (plus souvent que 'rencontré' *stricto sensu*) aux écoles d'été européenne d'histoire de la pensée économique, et qui ont, à différents égards, influencé ce travail de thèse, est trop grande pour que ses membres soient tous remerciés individuellement ici. Je m'en tiendrais à quatre de ces individus (pour les autres, cf. le premier paragraphe ci-dessus). Merci André Lapidus pour les perspectives toujours bienveillantes, et le scepticisme toujours terriblement bien placé, que vous avez porté sur mes travaux. Merci Guilhem Lecouteux pour toutes ces conversations sur l'économie comportementale, le paternalisme, mais surtout sur la théorie des jeux. Merci Paul Fourchard pour ces belles tranches de vie. Et merci Yaël Dosquet pour toutes ces profondes conversations, dans lesquelles tout se plie, se déplie, se replie, pendant que nous, on se remplit; je suis certain de te devoir plus d'une intuition, combien exactement et lesquels je ne sais pas trop (j'en identifie clairement au moins une).

De manière un peu moins organisée, il y a encore quelques chercheurs que je souhaite remercier pour leur influence comme pour leur bienveillance. En premier lieu, Philippe Mongin : merci pour ces quelques conversations dans lesquelles vous avez, comme dans tous vos articles cités ci-dessus, clarifié un ensemble de questions que je me posais, avec ce mélange de pertinence, de rigueur et de profondeur qui me fascine au plus haut point. Antoinette Baujard : merci, entre autres, mais surtout, pour cette lecture au scalpel de la première version de l'introduction générale et du premier chapitre. Cyril Hédoin : merci pour toutes ces relectures et ces conversations sur l'économie et la philosophie, c'est toujours extrêmement enrichissant. Ivan Moscati : merci pour ces conversations sur la théorie de l'utilité, mais surtout merci pour cette franchise que tu as pu avoir sur certaines dimensions de mon travail (notamment à Toronto). David Teira : merci pour cette conversation passionnante sur le rapport entre économie et philosophie et ces encouragements (c'est du moins comme cela que je l'ai pris) concernant la thèse de l'enchevêtrement. John Latsis : merci pour tes emails totalement inattendus, ça me touche beaucoup. Danielle Zwarthoed : merci pour toutes ces relectures et ces conversations passionnantes sur Amartya Sen et la philosophie. Floris Heukelom : merci, professionnellement, d'avoir entamé ce travail sur l'économie comportementale qui a été déterminant dans l'orientation de mon sujet de recherche, et, personnellement, pour tes diverses relectures toujours bénéfiques. Laurie Bréban et Hela Maafi: merci pour cette discussion qui a considérablement modifié des éléments de cette thèse Et enfin, Raphaël Giraud : merci énormément pour cette journée entière que tu as prise il y a deux ans pour reprendre ligne par ligne mon premier brouillon sur les effets de cadrage et la théorie de l'utilité, cette journée est restée une source de motivation et d'inspiration immense pour le travail de cette thèse.

Je souhaite également adresser une série de remerciements à cette spécialité locale moins connue des touristes que la pissaladière ou la socca: la sympathie des chercheurs niçois. Tout d'abord au GREDEG, je voudrais remercier Jacques-Laurent Ravix et Patrick Musso en leurs qualités de directeurs du laboratoire, pour les opportunités de recherche qui m'ont été permises grâce à eux. Pour leur assistance et leur bonne humeur dans la réalisation concrète de ces opportunités, je remercie également Laurence Gervasoni, Thérèse Marco, Agnès Moreau et Martine Naulet (dans le même esprit, je remercie chaleureusement Elisabeth Gazano). Pour les mêmes raisons, mais aussi pour beaucoup, beaucoup, d'autres, je remercie avec une affection toute particulière Muriel Dal-Pont Legrand : tu gères ! (Dans tous les sens du terme). Il y a quelques chercheurs que je souhaite remercier pour les moments en tête à tête qu'il mont accordé et qui ont influencé le fond comme la forme de mon travail de manière significative : Dominique Torre, Patrice Bougette, Flora Bellone, Joël Thomas Ravix et Aymeric Lardon. Je remercie également Christian Longhi et Adel Ben Youssef pour des mots particulièrement encourageants que vous m'avez séparément adressés après la première présentation orale du dernier chapitre de la thèse. Merci enfin à Alexandra Rufini et Sandye Gloria-Palermo d'avoir accepté, respectivement, de

discuter deux de mes textes particulièrement chaotiques. En ce qui concerne les doctorants, je remercie Guillaume Pupier pour, entre d'autres choses, m'avoir permis de ne pas entamer ma vie niçoise dans un vide social total; merci aussi aux autres pour diverses conversations intéressantes sur la microéconomie : Margot Ogonowska, Ankinée Kirakozian, Nabil Bennasr, Nabila Arfaoui, et particulièrement Anaïs Carlin. Pour tous les autres, cf. le premier paragraphe ci-dessus.

Dans une perspective plus interdisciplinaire, je souhaite remercier Laetitia Marcucci pour m'avoir donné l'opportunité de présenter mon travail à plusieurs reprises auprès des doctorants philosophes niçois, parmi lesquels je remercie Pierre Goldstein pour des conversations très éclairantes sur l'éthique et la métaéthique.

Et dans une perspective encore plus interdisciplinaire, je remercie tous les gens qui font de la MSH ce qu'elle est, c'est-à-dire non seulement une véritable maison, mais aussi un lieu de conversation interdisciplinaire que j'ai énormément apprécié. Tout d'abord merci Tobias Sheer, à la fois pour les opportunités de recherche que tu permets en ta qualité de directeur, mais aussi pour des conversations qui m'ont été extrêmement bénéfiques concernant les sciences cognitives, le langage, et même la métaphysique. Je remercie très sincèrement tout le personnel de la MSH pour leur disponibilité et leur bonne humeur au quotidien, Jean-Charles Briquet-Laugier, Françoise Beytet, Sylvie Grenard et Claire Gaugain, vraiment, merci beaucoup. Je remercie les chercheurs de BCL pour m'avoir proposé au tout début, quand je ne connaissais personne, de manger avec vous le midi, et ainsi avoir mes premières conversations avec des psychologues (merci Bruni, Morgan, Frédéric, Charlotte, Lucile), mais aussi avec des linguistes (merci Albert, Samir, Jonathan, Sylvain, Philippe). Je remercie tout particulièrement Damon Mayaffre et Richard Faure pour des conversations très instructives sur la place de la pragmatique en linguistique et en science sociales, Mustapha Chekaf pour d'autres sur les liens entre psychologie cognitive, logique et mathématique, et mes premières voisines de bureau (après Fara, que je remercie également), Rosa Volpe et Belén Jiménez pour avoir initié ce groupe de lecture qui m'a beaucoup apporté. Je remercie ensuite les chercheurs du LAPCOS de m'avoir initiée aux subtilités des relations internes entre sous-disciplines de la psychologie et de ses relations externes avec d'autres sciences sociales ; pour vous remercier à l'image de ce joyeux bazar où il est difficile d'attacher un thème précis à une conversation: merci Tania, grand Alex, Pierre, Zaineb, Eva, Ophélie, Fernanda, Daniel, Céline, Stéphanie, Orianne, Monica, petit Alex, Hanane et Romain. Un peu dans le même esprit, je remercie Ariana et Magali pour des conversations très éclairantes sur les 'studies'. Du côté du CEPAM, je remercie tout d'abord Frédérique Bertoncello d'avoir permis une présentation du travail sous-jacent au dernier chapitre de la thèse devant un public très interdisciplinaire, et pour des conversations intéressantes sur la place de l'archéologie dans les sciences, sur ce dernier point tout comme (avec des débordements fréquents vers d'autres sujets) Léa, Louise, Gaëlle, Liliane, Erwan, Emilie, Jean-Victor, Isabelle, Antonin, Eugénie, Alain, Arnaud, Benjamin, Martine et Gilles. Enfin, à l'intersection du GREDEG et de la MSH, je remercie Elise, Jamal, Loubna et Savéria pour l'ambiance de travail extrêmement stimulante qui règne dans le bureau 114 (et aussi pour avoir essayé de me faire comprendre quelque chose à la gestion), et Manuel, Bernard et Catherine pour des conversations passionnantes sur l'importance de la conversation.

Et puis pour toute une série de moments mémorables durant cette thèse, Sven et moi-même remercions (un peu dans le désordre): bien évidemment Svétlana, Tania et Savéria (les cooloques, j'en aurais bien écrit des pages et des pages sur vous, mais on va se calmer hein !), Mous (seau chaud col Ah ! qu'ils sont drôles tes jeux de mots), Ismaël 'la machine de guerre', Manuel, Tom (cap sur Babylone), Marie-Prune, Quentin, Dino, Guillaume, Lauren, Guilhem, Paul, Edouard, Yaël, Matthieu (bad boy, haha), Judith (ma tête dans le four didith), Cléo et Laure (c'était pas moi, c'était Sven), Erich, Niels, Francesco, Jean-Seb, Annie, Muriel, Agnès, Pierre, L'enfant sauvage, Fred, Céline, Vincent (on part quand ?), Pierrick, Robin,

Anaïs, Raphaël, Benjamin, Patrice, Margot, Ankinée, Magali, Sihem, Jazz-Jack-Patrick, Patrick Swayze, PsyKanopé, Nicolas, Marine, Slim (mon poète de la rue préféré), Alex, Aurélien, Alex, Pierre, Marine, Hanane, Romain, Louise, Benjamin, Alain, Claire, Jonathan, Morgan, Caroline, Camille ; avec quand même une petite pensée particulière pour les copains et les copines que j'aurais voulu voir plus souvent, mais la thèse à prise un peu trop le dessus (encore une fois un peu dans le désordre) : Léo a.k.a 'Double L' (ce bon vieux Mike à jouer un rôle à un moment donné de la thèse... il est nécessaire d'en parler), Léo a.k.a 'Duc', Torchon, Ju (merci pour tout pleins d'autres choses), Mireille (pareil), Valoche (pareil aussi !), Xav (pour ma soutenance je pense venir en girafe), Etienne (internet !) et Victoria a.k.a 'BigeToz', Quentin et Charlotte a.k.a Thierry et Marie-Chantal, Karl (il va vite il va vite), Guillaume Lewis ne perd jamais, Vil, Max, Justine, Photographe, et bien évidemment Ghislain Balagna (pour ma soutenance je pense soutenir à environ 130dB, et ne t'inquiètes pas pour le soir, j'en placerai une pour Pasc').

Enfin, je remercie ma famille sans qui rien de tout cela n'aurait était possible : merci Maman et Papa pour l'amour, la curiosité le goût du travail bien fait, et puis Ciacia pour être toujours à la cool ; et puis pour les même raisons, merci Eliane, Christelle, Mich, Kink, Raq, Juju, Tom, Valentin, Eddie, Ondine, Lydie, Karine, Laurène, et bien évidemment Mémé Dédé, Mémé Mado... et Pépé Co : les pages en haut, au final, « c'est pour dire ».

Abstract

L'économie comportementale et le rôle du langage dans la rationalité économique: une perspective méthodologique.

Dans cette thèse, nous proposons une perspective méthodologique sur le double rôle du langage dans la rationalité économique, les utilisations de langage par les économistes pour la théoriser et les utilisations de langage par les agents économiques pour l'exprimer, pour clarifier trois principales questions (et leurs connexions) qui sous-tendent les débats entre économie comportementale et économie standard : le problème de l'unification théorique vis-à-vis des trois dimensions de la rationalité économique, la question de l'interdisciplinarité entre économie et *P*sychologie, et le problème du positif/normatif dans les modèles de comportements individuels. Concernant le problème du positif/normatif et le rôle du langage dans les comportements des agents économiques, notre intention est de fournir, au-delà de la simple clarification, une critique constructive des contributions de l'économie standard comme de l'économie comportementale. Suivant la position de l'enchevêtrement du philosophe Hilary Putnam et des philosophes-économistes Vivian Walsh et Amartya Sen, il est soutenu que l'économie tant standard que comportementale propose une articulation insatisfaisante des dimensions positive et normative dans les modèles de comportements individuels; et que la reconnaissance de l'enchevêtrement de faits, de valeurs et de conventions peut être théoriquement et empiriquement fructueuse. Prêter attention au rôle du langage dans les comportements des agents économiques montre *parfois* qu'un comportement apparemment irrationnel peut en fait être défendu comme rationnel; c'est pourquoi nous soutenons que, et montrons comment, l'axiome implicite - connu sous le nom d'invariance à la description - dans les modèles standards de comportements individuels empêchant l'influence du langage doit être affaibli (mais pas complètement supprimé), contrairement aux positions de la plupart des économistes standards et comportementaux.
Mots-clés: économie comportementale, théorie de la décision, effet de cadrage, renversements de préférence, rationalité, méthodologie de l'économie, philosophie de l'économie

A methodological perspective on behavioral economics and the role of language in economic rationality.

In this dissertation, we propose a methodological perspective on the twofold role of language in economic rationality, economists' uses of language to theorize it and economic agent's uses of language to express it, can clarify three main issues (and their connections), underlying the behavioral *versus* standard economics debates: the issue of the theoretical unification regarding the three dimensions of economic rationality, the issue of interdisciplinarity between economics and *P*sychology and the positive/normative issue within models of individual behaviors. Regarding the positive/normative issue and the role of language in the behaviors of economic agents, the intention is to provide a constructive criticism of contributions from behavioral as well as standard economists. Following the entanglement thesis of philosopher Hilary Puntam and philosophers-economists Vivian Walsh and Amartya Sen, it is argued that both standard and behavioral economists propose an unsatisfying articulation between the positive and normative dimensions of models of individual behaviors; and that recognizing the entanglement of facts, values and conventions can actually be theoretically and empirically fruitful. Paying some attention to the role of language in the behaviors of economic agents may *sometimes* show that a seemingly irrational behavior can in fact be defended as rational; hence we argue that, and show how, the implicit axiom – known as 'description invariance' – in standard models of individual behaviors preventing the influence of language needs to be weakened (though not dropped entirely), contrary to the positions of most behavioral and standard economists.
Keywords: behavioral economics, decision theory, framing effect, preference reversals, rationality, economic methodology, philosophy of economics