Aix-Marseille Université

École Doctorale de Sciences Économiques et de Gestion (N372)

Faculté d'Économie et de Gestion

Groupe de Recherche en Économie Quantitative d'Aix-Marseille (GREQAM)

Aix-Marseille School of Economics (AMSE)

Thèse de Doctorat de Sciences Économiques

Présentée par

João V. FERREIRA

en vue de l'obtention du grade universitaire de docteur

Conflicted Individuals:

Essays on the Behavioral Implications of Multiple Preferences

Soutenue publiquement le 02/10/2017 devant le jury composé de :

Arthur SCHRAM	University of Amsterdam and European University Institute, *Rapporteur*
Robert SUGDEN	University of East Anglia, *Rapporteur*
Salvador BARBERÀ	Autonomous University of Barcelona and Barcelona GSE, *Examinateur*
Alain TRANNOY	Aix-Marseille Université, *President du Jury*
Nicolas GRAVEL	Aix-Marseille Université, *Directeur de thèse*

To my (three) parents
and to Graça Fernandes,

Cada um de nós é vários, é muitos, é uma prolixidade de si mesmos. Por isso aquele que despreza o ambiente não é o mesmo que dele se alegra ou padece. Na vasta colónia do nosso ser há gente de muitas espécies, pensando e sentindo diferentemente.

Chacun de nous est plusieurs à soi tout seul, est nombreux, est une prolifération de soi-mêmes. C'est pourquoi l'être qui dédaigne l'air ambiant n'est pas le même que celui qui le savoure ou qui en souffre. Il y a des gens d'espèces bien différentes dans la vaste colonie de notre être, qui pensent et sentent différemment.

Each of us is several, is many, is a profusion of selves. So that the self who disdains his surroundings is not the same as the self who suffers or takes joy in them. In the vast colony of our being there are many species of people who think and feel in different ways.

Bernardo Soares (Fernando Pessoa), O Livro do Desassossego

Abstract

In this thesis I explore decision making models based on multiple preferences. In the first part of the thesis, I analyze some of the implications of adopting multiple preferences in economics and different ways in which they can be conceptualized and used within this field. In particular, I review some of the positive and normative consequences of preferences over preferences (Chapter 1), the behavioral (in)distinguishability of the single and multiple preferences models (Chapter 2), and introduce a new framework of choice with time in which models of changing preferences can be more easily characterized (Chapter 3). The second part of the thesis is devoted to the theoretical and empirical analysis of economic meaningful behavior that can be represented as if it is the result of decision making with multiple preferences. In particular, I build a model to study the effects of multiple preferences to political behavior (Chapter 4), and run an experimental study to distinguish different motivations behind a potential intrinsic value of holding a decision right (Chapter 5).

Keywords: Multiple preferences; Revealed preference theory; Reflexive preferences; Preference change; Behavioral welfare economics; Time; Spatial voting; Conflicted voters; Intrinsic value; Decision rights; Cross-cultural experiment.

JEL classification: B4; C91; D01; D03; D6; D7; P16.

Résumé

Dans cette thèse, j'explore les modèles de prise de décision basés sur des préférences multiples. Dans la première partie de la thèse, j'analyse certaines des implications de l'adoption des préférences multiples en économie et de différentes façons dont elles peuvent être conceptualisées et utilisées dans ce domaine. En particulier, je révise certaines des conséquences positives et normatives des préférences sur des préférences (chapitre 1), la distinction comportementale entre des modèles de préférences uniques et des modèles de préférences multiples (chapitre 2), et j'introduis un nouveau cadre de choix avec le temps dans lequel les modèles de préférences multiples peuvent être plus facilement caractérisés (chapitre 3). La deuxième partie de la thèse est consacrée à l'analyse théorique et empirique du comportement économique qui peut être représenté comme s'il résulte de la prise de décision avec des préférences multiples. En particulier, je construis un modèle pour étudier les effets des préférences multiples sur le comportement politique (chapitre 4) et je mène une étude expérimentale pour distinguer les différentes motivations derrière une potentielle valeur intrinsèque du droit de décision (chapitre 5).

Mots-Clés: Préférences multiples; Théorie des préférences révélées; Préférences réflexives; Changement de préférences; Économie comportementale du bien-être; Temps; Vote spatial; Électeurs en conflit; Valeur intrinsèque; Droits de décision; Expérience interculturelle.

Classification JEL: B4; C91; D01; D03; D6; D7; P16.

Acknowledgments

According to Harry G. Frankfurt, one of the keys to a meaningful and fulfilled life is to pursue wholeheartedly what one cares about. Four years have passed since I started this dissertation, and today I am convinced that this endeavor was certainly a means to many ends that I wish to pursue wholeheartedly. Meanwhile, the experience of writing it was not always wholehearted. It was not always easy to find purpose, and the relationship between the time consumed and the resulting product was not always self-evident. That is one of the main reasons why having met so many good, interesting, and inspiring people, and having had the support of so many others was so important.

I wish first to acknowledge the constant and invaluable support of my supervisor, Nicolas Gravel. We have met when I came to Aix-Marseille as a second-year master student, five years ago, and since then he has provided me guidance and encouragement. He has been always available, and I have enjoyed the many discussions we had during these 4+1 years. The shape of the chapters that compose this dissertation is very much influenced by his comments, remarks, corrections, suggestions, and, I guess, his general view about these topics. It was also a pleasure to work with him on the projects we have started together, and to arrive to what is now Chapter 3 of this dissertation. You have also been kind to me in many social occasions, and understanding in non-academic matters. I have learned invaluable lessons with you, and I admire you as a person and researcher. I am deeply grateful for all of this.

I am also indebted to my remaining co-authors, with whom I have learned a great deal during these last four years. Sacha-Bourgeois Gironde has since my second year

of master encouraged me to develop my ideas, always giving an original outlook at them. Fruit of our joint work is now Chapter 4 of this dissertation. More recently, collaborating with Nobuyuki Hanaki and Benoît Tarroux in Chapter 5 of this dissertation has been a smooth and enriching experience. I have benefited from their expertise in experimental economics, and learned continuously with our discussions about the different options that we had to make. They have also been very supportive concerning my future and in general academic matters, and I am very grateful for that.

I am also very grateful to the members of the jury, and honored that they have accepted to make part of it. I wish to start with Alain Trannoy, that besides president of the jury, has been present in many steps since I have started writing this dissertation. He has attended many of my presentations, and always made valuable and thoughtful suggestions. As a member of the jury, he made again significant comments that helped me to shape the last version of this dissertation. You had also thoughtful and encouraging words about the future, and I wish to thank you for that.

Arthur Schram and Robert Sugden, the two referees of this dissertation, made valuable and extensive comments about the exposition and content of the manuscript that definitely changed its final shape. I am very grateful for your thoughtful reading of the dissertation, and for having raised important concerns about its content that I have tried to offset, thanks to you, in the last version of the manuscript. I have had the opportunity to meet Arthur Schram as a teacher at a crash course in experimental economics, which was undoubtedly an enriching experience. You were also kind to me when we had the opportunity to discuss, and I hope we have similar opportunities in the future. Though I have not yet had the opportunity to meet Robert Sugden in person, I also hope we have such opportunity in the future.

Last, but not least, Salvador Barberà has made many valuable comments upon my research. In particular, you made me look - and become curious - about literatures that I was less familiar. You have also been very kind to me in the occasions we have met, and I have enjoyed the interesting and broad discussions we had. As with the other members of the jury, I hope such opportunities arrive again.

Many other professors and/or researchers have been sources of support and inspiration during these four years. I wish first to thank Marc Fleurbaey for having welcome me at Princeton during three months of the second year of writing this dissertation. He was very attentive and available in a regular basis, and I have certainly learned a lot with our discussions. Though much of what we have discussed is not directly linked to the final content of this dissertation, I have widened my interests and deepen my critical reasoning thanks to you. You also gave me glimpses of how to pursue meaning in research, that I take with me since then. Being at Princeton was a very enriching experience at many levels, and I wish to thank you for allowing me that opportunity.

I wish also to thank Miriam Teschl and Stéphane Luchini for the many occasions in which we had enriching debates, and for the many interdisciplinary seminars and workshops that widened my perception and interests. I am also indebted to the valuable comments and suggestions you made about my research. You have also been very kind to me outside the workplace, and I hope we continue to have opportunities to meet, discuss, and "enjoy the good life". I am also grateful to Sebastian Bervoets, who since he was my professor in the master 2 of economic philosophy and referee of my master thesis, that he has been present, supportive, and often gave me valuable suggestions about my research. I wish also to thank Antonin Macé, for his kindness and extensive comments upon a previous version of Chapter 4. I have also benefited from the support and/or thoughtful discussions with so many other people over the years, such as Mathieu Faure, Nick Sheard, Maria Bigoni, Erik Schockaert, Hubert Stahn, Raouf Boucekkine, Federico Trionfetti, Thibault Gadjos, Marc Sangnier, Andreas Schmidt, Roberto Iacono, Joel Anderson, Claudio Zoli, Jeroen van de Ven, Fabrice Le Lec, Nicolas Jacquemet, Geoffrey Brennan, just to name a few.

There are also some that made this possible just before it started, and others that eased so much the process while it lasted. As for the former, I am thinking about José Tavares, João Cotter Salvado, Klaus Desmet, Alain Leroux, Claude Gamel, Hugo Domic, that either helped me or inspired me to learn. As for the latter, I am thinking

about Bernadette, Isabelle, Agnès, Corinne, Aziza, Mathilde, Kaina, Yves, Gregory, Marine, and Gérald. Thank you for your precious help and good humor, even when facing someone not always well organized as myself.

I have been lucky to be part of a dynamic group of Ph.D. students. Many of them have shared this ride with me, and I can only thank them for their support and their friendship. Clémentine, who is the one that I know for the longest, and that was a great friend all along. I am grateful for your support either in personal or in academic matters, that often helped to keep me on track. Justine and Nicolas, my French parents and neighbors at the Panier. I could not wish a better couple to welcome me to France. Maxime, with whom we have met so many times, we have made Marseille a mixed *terroir* of debate and good fun. Damien, with that sparkle of craziness that I love. Thank you for your friendship, your good faith, your humor, and your mathematical deviations. Ugo, the Italian, maybe stallion, but foremost kind and good friend. Thank you for being present, and for having that curiosity that made us discuss so many projects together, and arrive to one in which we find meaning. Majda, a lively source of good energy. Lara, with whom was a pleasure to grow. Laure, that so often brought a bit of nature to the office. Guzman, who was always a very good surprise. Tanguy, Stéphane, Alberto, with whom I shared many good moments these last two years. Ilia, Edwin, and Audrey, that made the end of this dissertation so much happier. And surely many others: Edward, Victor, Nicolas (the many), Houda, Kadija, Régis, Florent, Adrien, Anwesha, Laila, Khalid, Lise, Gilles, Karine, Anastasia, Marion, Emma, Vivien, Manel, Samuel, Solène, Pauline, Victorien, Cyril, Océane, Etienne, Vera, Yannick, Yezid, Armel, Estefania, Andrea, Marie-Christine, François. I wish to thank you all for creating a lively and friendly place to work, and for all those moments that are now good memories. This dissertation would have definitely be more difficult to accomplish without your friendship.

I would also like to thank the Juliens from Princeton, Kabira, Trygve, and Laudine for their friendship and valuable discussions about topics that relate with the content

of this dissertation. Ilia, Edwin, and Laure for their help with the translation from English to French of the abstract and general introduction of this dissertation.

These acknowledgments would not be complete without a word for the people and "places" that made Marseille the city *she* became for me. Michèle, Silvain, Philemon, Johanna, Perrine, Natalia, Raphael, Lionel, Laurent, Maria, Esther, Sonia, Fred, Haik, Elise, Arsène, Dianette, Eric, Amandine, Alex, Pia, Julia, Hanna, Mandy, Victor, Audrey, Rocio, Charlotte, Benoît, Thomas, Stéphane, *le manolo*, *la bagagerie*, *le dojo*... You have brought joy and meaning to so many of my days. Thank you!

Finally, there is the people that, as a consequence of this endeavor, I have seen less often that I would like to. Many are the friends that I miss, and that bring me joy and support whenever I go "home". Among these, there are some that during these four years were not only *there* to receive me but were also constant and invaluable sources of support at the other end of the screen. Miguel, the most faithful rock 'n' roller. Rogério, the all-encompassing friend. Diogo, the brother of another mother. You were like three musketeers during these four years (me not having neither the style nor the skills of D'Artagnan). I am very fortunate to have you as friends.

I wish also to thank my brother and sister, for being examples in the path of learning. I am also grateful for the love with which you have received me whenever I have come back to Portugal. Thank you also to Filipa and Arnau and all my beautiful nieces for bringing me joy whenever I visit.

I am very grateful to Graça, who is someone who I truly admire for her courage, wisdom, and grace. In the past few years, in which we had the possibility to discuss at a deeper level, you have been a bright light in all the occasions we have met. You have inspired me to learn and become a better person. I only wish to arrive at your age with some of your grace. Thank you for being such an inspiration.

All in all, I consider myself an extremely fortunate person. One of the main reasons is that I could not wish more supportive and caring parents. As usual, they were there all the way. I wish to thank my aunt, my second mother, for her unconditional love. And *Pai*, *Mãe*, thank you for being as you are. I am very fortunate to have grown

up watching you, your relationship, and to continue to have you by my side. You are examples to me in many ways. I can only thank you for the way you have always supported me, for the mindful and tolerant education that you gave me; certainly all this would not have been possible, or at least not as easily possible, without you.

Lastly, I wish to thank Anne, who was the person most present during these four years. I am grateful for your patience, caring, love, and support. This experience is intimately linked with you, and I have learned as much with you, with us, as with the research itself. I am grateful for all the moments we have shared, and for all the memories that will last. I am also grateful for having learned, more recently, that it is possible to love the same person with many selves. I can only thank you for all of that.

In his dissertation, Henry Bergson wrote that "language is not meant to convey all the delicate shades of inner states". I hope, nonetheless, to have conveyed a glimpse of what mine might have been during these four years. *Obrigado*!

Mainly written on the 24th of July 2017,
between Marseille and Lisbon.

Contents

List of Tables

List of Figures

General Introduction

The aim of this introduction is first to motivate the research undertaken in this thesis (Section 0.0.1), second to conceptualize the notion of multiple preferences (Section 0.0.2), third to provide an overview of the research chapters (Section 0.0.3), fourth to briefly discuss the methods and the epistemological view adopted in this research (Section 0.0.4), and finally to introduce some general notation that will be used the first part of the thesis (Section 0.0.5). Some of the related literature is discussed along this introduction, namely in Section 0.0.2. The list of references is provided at the end of the introduction.

0.0.1 Motivation

Fernando Pessoa, one of the most prolific of Portuguese writers, and in my and many others' view the most brilliant of them, wrote under the name of several fictional figures that he had created. These figures were, according to him, more than pseudonymous. They were, instead, his "heteronyms", endowed with their own biographies, appearances, feelings, and worldviews. They wrote better or worse Portuguese than one another, about different topics, and in different styles.

If most of us do not recognize such independent figures in ourselves, the idea that people are multi-faceted has a long tradition in philosophical thought. One can trace back at least to Plato the idea that human beings are internally divided. According to Plato, the clash between moral reasoning and human immoral passions was central:

> "First the charioteer of the human soul [reason] drives a pair, and secondly one of the horses is noble and of noble breed [moral], but the other quite the opposite in breed and character [passions]."
>
> Plato, Phaedrus

Conflicting motivations continued to be a topic in philosophical thought for the centuries to follow. For example, there is considerable written evidence that by the seventeenth and eighteenth centuries authors focused on the conflict between morality, the human immoral passions, and the pursuit of material interest that, until then, was depreciated itself as the immoral passion of avarice (see Hirschman 1977). But it was only on the nineteenth century that multiple preferences, in the form of multiple identities or selves, became a subject of study. William James first conceptualized the notion of "multiple selves" as follows:

> "Properly speaking, *a man has as many social selves as there are individuals who recognize him* and carry an image of him in their mind. To wound any one of these his images is to wound him. But as the individuals who carry the images fall naturally into classes, we may practically say that he has as many different social selves as there are distinct *groups* of persons about whose opinion he cares. He generally shows a different side of himself to each of these different groups. Many a youth who is demure enough before his parents and teachers, swears and swaggers like a pirate among his "tough" young friends. We do not show ourselves to our children as to our club-companions, to our customers as to the laborers we employ, to our own masters and employers as to our intimate friends. From this there results what practically is a division of the man into several selves; and this may be a discordant splitting, as where one is afraid to let one set of his acquaintances know him as he is elsewhere; or it may be a perfectly harmonious division of labor, as where one tender to his children is stern to the soldiers or prisoners under his command."

William James, The Principles of Psychology

Nowadays, psychology and behavioral economics have provided considerable empirical evidence suggesting that behavior is often due to or can be explained by conflicting motivations, multiple identities, or the different roles that people lead in their lives. For example, some experiments suggest that priming one of two identities (the Asian or the American identity of Asian-American subjects) triggers different behavioral responses in terms of patience (Benjamin, Choi and Strickland 2010) and cooperation (LeBoeuf, Shafir and Bayuk 2010). Similarly, some experiments suggest that primes of intelligence-related social constructs such as "a professor" or "Albert Einstein" (as opposed to "a supermodel" or "Claudia Schiffer") affect self-perceived intelligence, one's self-concept, and subsequent behavior in terms of test scores (Dijksterhuis, Spears, Postmes, Stapel, Koomen, van Knippenberg and Scheepers 1998; Schubert and Hafner 2003; LeBouef and Estes 2004).[1] In addition, accumulating evidence as well as casual observation and introspection indicate that choice behavior is often the result of *endogenous preferences*, i.e., preferences that dependent upon the experience of the decision maker.

However, the dominant neoclassical approach in economics is to summarize the individual tastes, values, interests, and goals into a single, stable, and exogenous preference relation. According to this view, an individual's personal identity cannot change according to the context or over time. There is no internal conflict that an individual is not able to resolve, and no evolution underlying his experience trough time.

Choice behavior is, in the neoclassical choice model, assumed to result from the maximization of this stable and exogenous preference relation. The observational implications of this model are described by the traditional revealed preference axioms, such as the Weak and Strong Axioms of Revealed Preference (see Sen 1971 for

[1] See Wheeler, DeMarree and Petty (2007) for a review of the evidence on the effects of primes on self-concept change and subsequent behavioral change. See e.g. Marks and MacDermid (1996) for an essay on multiple roles and their effect on subjective measures of well-being.

a review). Under some conditions, these axioms are necessary and sufficient to describe a set of choices *as if* resulting from the maximization of a stable and exogenous preference.

This approach is problematic for at least two reasons. First, it may be deficient in terms of description and prediction of economic behavior. In particular, the rationality requirements that the neoclassical choice model demands are not consistent with patterns of behavior due to changing preferences, several learning models, and other contextual and social determinants of behavior. Second, it may lead economists astray in terms of welfare inference and welfare ranking of different social states. For one thing, preferences and choices often fail to reveal individuals' well-being, since they can, among other things, be the result of cognitive dissonance, a blatant mistake, or manipulation. But it is also the case, as it will be argued in Chapter 1, that the single preference model eschews normatively relevant information on what people *value*, what they *care* about, and *who* they wish to be or become.

An alternative view to the traditional rational choice model is to assume that the economic agent is driven by multiple preferences. According to this view, choice behavior is not the result of the maximization of a single preference but instead the result of the aggregation, conflict, or the change between multiple preference relations. Many of the decision making models that are of interest to economists and that the standard theory cannot explain, including changing preferences and preference formation, are due to or can be explained by multiple preferences, identities, or selves. Similarly, the evolution of individuals' preferences according to their experiences can be due to or can be explained by multiple preferences over time.

The aim of this thesis is to explore models of decision making based on multiple preferences as an option to the single preference paradigm. In the first part of the thesis, I explore some of the (behavioral) implications of adopting multiple preferences in economics. I review some of the positive and normative consequences of this proposal (Chapter 1), the behavioral distinction between the single and multiple preferences models (Chapter 2), and introduce a new framework of choice with time

in which models of changing preferences can be more easily characterized (Chapter 3). The second part of the thesis is devoted to the theoretical and empirical analysis of economic meaningful behavior that can be represented as if it is the result of decision making with multiple preferences. In particular, I build a model to study the effects of multiple preferences to political behavior (Chapter 4), and run an experimental study to distinguish different motivations behind a potential intrinsic value of holding a decision right (Chapter 5).

Before proceeding to an overview of the research chapters, I discuss and conceptualize the notion of multiple preferences and provide a taxonomy of multiple preferences models that may be useful to contextualize the research carried out in this thesis and to indicate future avenues of research.

0.0.2 Multiple Preferences

There is by now many decision making models based on multiple preferences that are used to explain economic behavior. One example is given by the dual-process or dual-system theories that are now prominent in economics (Thaler and Shefrin 1981, Bernheim and Rangel 2004, Fudenberg and Levine 2006, among many others).[2] The central assumption shared by all these decision making models is that some economic behavior is the result of the interplay of two broad types of decision making, one based on reasoned/reflective deliberation and another on impulsive/automatic decisions. These models are used to explain relevant economic behavior such as addiction (Bernheim and Rangel 2004) and intertemporal choice (Thaler and Shefrin 1981; Fudenberg and Levine 2006).

Preferences, Self, and Identity. Since the notions of multiple preferences, multiple selves, and multiple identities are often used and sometimes interchanged in the literature it is useful to precise what I mean by them at this point. I use the term *multiple preferences* as an "umbrella" concept, that encompasses a collection of order-

[2] See Alós-Ferrer and Strack (2014) for a review.

ings (based e.g. on different motivations, cares, or points of view), multiple selves, or identities. I consider that preferences are or can be seen as an expression of (one)self, and that different decision making models point towards different underlying notions of the personal identity of the economic agent.[3] I use the term *multiple selves* to refer to the cases where the multiple preferences are modeled as " subagents" that interact with each other as if they were players in an interpersonal game. Finally, I use the term *multiple identities* to refer to the different (social) identifications that individuals may hold for different groups or adopt in different contexts.

A Tentative Taxonomy of Multiple Preferences Models. Decision making models based on multiple preferences can be distinguished according to many criteria. Based on my previous discussion and other reviews (e.g. Ambrus and Rozen 2013), some plausible criteria to differentiate multiple preferences models are: (i) whether or not all preferences are active at each period[4], (ii) whether or not these preferences are stable over time, (iii) and if the multiple preferences are independent or commensurable into a single preference at each point in time. The first criterion distinguishes the models that take behavior at each period as the result of the maximization (or other process) of one of multiple preferences, from the ones that model behavior as the result of the aggregation (or other process) of multiple preferences at each period. The second criterion distinguishes the models that assume stable preferences from those that assume no *a-priori* restriction in terms of temporal consistency. Finally, the third criterion distinguishes the models that assume a single ranking of alternatives at each period from the ones that model the different preferences as independent orderings, selves, or identities.

Table 5.1 presents a tentative taxonomy of different multiple preferences models that emerges from the intersection of these three criteria. In what follows I discuss

[3]Though, certainly, I do not consider that preferences exhaust an individual's identity or self. Similarly, an individual's personal identity is also often a part (even if a considerable one) of one's self-concept. Understanding these connections allow, among other things, to distinguish different underlying views over the individuals' mode of reasoning and to differentiate distinct views over a person's well-being and responsibility. See Oyserman, Elmore and Smith (2012) for more on the connection between the notions of the self, self-concept, and identity.

[4]Ambrus and Rozen (2013) differentiate multiple preferences models based on this dimension.

each one of these representations of the economic agent with a brief relation to the literature, with the exception of the *static preference* that is nothing more than the traditional rational choice model.

Table 1 – Multiple Preferences Models

	Stable	**Not Stable**
Single preference	Static Preference	Evolving Preferences
All preferences active	Simultaneous Preferences	Successive Preferences
One (of many) active	Alternating Preferences	

Evolving Preferences. This representation conceptualizes the economic agent as if endowed with one personal identity that evolves over time. This represents the individual as an evolving agent that makes her decisions according to an (endogenous) sequence of multiple preferences.

In economics, most models consistent with this view assume an exogenous sequence of preferences. For instance, Gul and Pesendorfer (2001, 2004, 2005) model of intertemporal choice and the changing preferences model that is characterized in Chapter 3 are consistent with the exogenous evolution of preferences. It is worth noting that psychology, philosophy, and neuroscience support the view that a person's identity evolves over time.[5]

As it is argued in Chapter 1, this process of evolution may be mediated and/or represented by preferences over preferences, also known as hierarchical preferences, meta-preferences or second-order preferences.[6] Some authors have advocated the use of preferences over preferences in economics (e.g. Sen 1977; Hirschman 1984), and in Chapter 1 I sketch two hierarchical models that could be used for both positive and normative economic analysis. The representation (or not) of hierarchical preferences is yet another meaningful distinction between different multiple preferences models.

[5]See e.g. Gallagher (2000) for a brief review of some theoretical and empirical arguments in favor of a narrative (evolving) identity.

[6]See Frankfurt (1971) for the philosophical basis of this notion. The term hierarchical preferences is often used in the literature to refer to agency models that include higher-order preferences (e.g. Elster 1985) and I use it here and in Chapter 1 for convenience, though it may be somewhat misleading. In particular, I take the hierarchy to be formal, instead of a reflection of some sort of dominance.

Simultaneous Preferences. This representation captures decision makers who have a collection of independent preferences that are stable and active in all periods. Models in this vein represent the economic agent as if endowed with a collection of simultaneous preferences, i.e., a plurality of distinguishable identities, motivations, points of view, or cares that are fixed over time.

For instance, Aizerman and Malishevski's (1981) *pseudo-rationalization* model, one of the first to provide observable properties of a model based on multiple preferences, belongs to this category. For every choice situation the agent selects the union of the maximal elements of all preferences, i.e., the elements that are "best" for at least one preference.[7] More recently, Cherepanov, Feddersen and Sandroni (2013) build on a similar representation to propose a testable model in which an agent uses a collection of preferences (interpreted as different stories that an agent tells herself) to *rationalize* a subset of options from which she can choose from. Any option is rationalizable in this sense if it is at least best for one of the agent's preferences (i.e., the rationalizable options are the union of the maximal elements of all preferences, the ones that were selected in Aizerman and Malishevski 1981). Then, for every choice situation the agent chooses, among these options, the one that is the most preferred (maximal) according to a single, stable, and exogenous preference relation. Models of choice by sequential or lexicographic procedures, such as Tversky (1969), Manzini and Mariotti (2007, 2012), and Apesteguia and Ballester (2013) also belong to this category. In these models, at all choice situations an arbitrary number of preference relations (rationales with different properties according to the model) is applied sequentially to single out one alternative to be chosen.

Successive Preferences. The difference of models based on successive preferences to those of simultaneous preferences is that for the former the collection of preferences is not necessarily stable across time. In decision making models based on successive preferences there may exists a new set of multiple active preferences at any given point in time.

[7] Eliaz, Richter and Rubinstein (2011) explore an analogous approach for two preferences.

For instance, many of the dual-self models of intertemporal choice assume that a long-run self interacts with successive short-run selves. Fudenberg and Levine (2006, 2012), for example, model a rational agent endowed with one stable and far-sighted self (a "planner") that interacts with a *new* myopic self (a "doer") at each period (Fudenberg and Levine 2006) or after every few periods (Fudenberg and Levine 2012).

Alternating Preferences. This representation conceptualizes the economic agent as if endowed with a collection of (stable or unstable) preferences and that she alternates between them from one period to the other.[8] The difference of this representation to that of simultaneous preferences is that one preference, and not two or more, is going to dictate the decision at each period.

For example, the *reason-based* theory developed by Dietrich and List (2013, 2016) is consistent with this representation. In their model, an agent is represented as a family of preference relations over all possible *motivational states*, defined as a subset of all possible *motivationally salient* properties of those alternatives (the properties that the agent focus on). Then, an agent alternates from one preference relation to another according to the motivational state in which she happens to be.[9] *Random preference* models, such as the ones of Becker, DeGroot and Marschak (1963), Barberà and Pattanaik (1986), McFadden and Richter (1990), Loomes and Sugden (1995), Gul and Pesendorfer (2006), Apesteguia, Ballester and Lu (2017), among others, may also be interpreted as if based on alternating preferences.[10] In these models, the individual preference that is active in a given choice situation is drawn at random from a pool of potential preferences.

The view that people often behave in a "single-minded" way despite being best seen as a collection of preferences is also shared, for instance, by Schelling (1984)

[8] Another interpretation consistent with this category would be a model with a *dictator self*, in which a unique and stable self (always the same among many) takes the decisions in all periods. However, this representation seems to be hardly interesting, since it is observationally equivalent to a static preference and, as pointed by Steedman and Krause (1985, 208), most people avoid being or becoming purely *one*-dimensional as such interpretation would presuppose.

[9] See also Tversky and Kahneman (1991) who use alternating preferences to model reference-dependent behavior. In their model, an agent has a fixed preference relation for each *reference state* that is broadly defined as the agent's current position.

[10] See Fishburn (1999) for an old but comprehensive review.

and Gigerenzer and Selten (2000). According to Schelling (1984) people are best represented as a collection of "values centers" that share the same beliefs and reasoning capacities but differ in terms of volitions. According to this view, one value center (or self) will act as if a dictator at each period, winning "the intimate contest for self-command" at that period (see Schelling 1984, 57-81). According to Gigerenzer and Selten (2000), cues in the environment will single out one of many heuristics to an agent. Since these heuristics advance a particular end, agents will act according to a single criteria and in a single-minded way at each choice situation.

An important aspect that distinguishes these models is that while evolving preferences models treat the agent as a *unity of agency* (as the traditional rational choice model based on a static preference), simultaneous, successive, and alternating preferences models represent the economic agent as a divided agent for whom several orderings, identities, or selves dispute the internal contest for self-command. This distinguishes two broad representations of the economic agent based on multiple preferences: (i) an **evolving agent** that decides according to the evolution of a single preference, and (ii) a **conflicted agent** that decides according to the conflict between multiple preferences. While there is by now an extensive economic literature on models based on conflicted agents that disaggregate a person's unit of agency into multiple orderings, identities, or selves, less effort seems to have been made to model an evolving agent that takes decisions based on a personal identity that changes according to her experience through time.

0.0.3 Overview

The research carried out in this thesis is divided in five chapters. I set the stage in Chapter 1, with an appraisal of some of the positive and normative consequences of adopting models based on multiple preferences in economics. I contrast this option against the traditional rational choice model and recent behavioral models that treat behavior inconsistent with the maximization of a stable preference as mis-

taken. I argue that, instead of avoiding "authentic" preference change, it is important to distinguish mistakes from inconsistent behavior that results from preferences (or preference change) that individuals identify with. These are cases of reflexive (self-authenticated) preferences or preference change even if they contradict the maximization of a stable and exogenous preference. I sketch two hierarchical models that represent some of these ideas, and discuss how they could relate with the conflicted agent and evolving agent models in order to represent reflexive and non-reflexive preference change. I argue that making the distinction between reflexive preferences/preference change and non-reflexive ones may lead to better description and prediction of economic behavior, and that collecting non-choice data on which preferences (or preference change) individuals identify with may be useful for normative economics, in particular as a refinement to welfare rankings currently used in behavioral welfare economics.

In Chapters 2 and 3 I bring this analysis to formal choice theory. The common interpretation given to choice behavior that satisfies the traditional revealed preference axioms is that it is the result of the maximization of a stable and exogenous preference relation. In Chapter 2, I show that choice data alone does not enable one to rule out the possibility that the choice behavior that satisfies the revealed preference axioms is instead the result of the aggregation of a collection of distinct preferences. In particular, I show that any ordering is observationally equivalent to a majoritarian aggregation of a collection of distinct dichotomous orderings. I also show that any ordering is observationally equivalent to a Borda's aggregation of a collection of distinct linear orderings. I use these two examples and related results to discuss observational "indistinguishability" and model selection. I argue that the issue of indistinguishability may extend to contexts where some choice behavior may be the result of either an individual or a collective decision; however, I argue as well that questions concerning the plausibility of the different explanatory models and if it is important to identify the underlying model of choice behavior need to be asked before considering theoretical indistinguishability problematic. In the case that indis-

tinguishability is indeed problematic, an open question remains about which methods - besides subjective non-choice data - should be used to identify the underlying model of choice behavior.

In Chapter 3, which is based on joint work with Nicolas Gravel, we introduce a framework for the analysis of choice behavior when the later explicitly depends upon time. We relate this framework to the traditional timeless choice-theoretic setting, and illustrate its usefulness by proposing three possible models of choice behavior in such a framework: (i) changing preferences, (ii) preference formation by trial-and-error, and (iii) choices with endogenous *status-quo* bias due to inertia in preferences. We provide a full characterization of each of these three choice models by means of revealed preference-like axioms that could not be formulated in a timeless setting. While only the first of these is rationalized by a model of decision making based on multiple preferences, our analysis is suggestive of the potential of this framework to study other choice models motivated by endogenous and multiple preferences.

Chapter 4 is devoted to the relation of multiple identities and political behavior, and is based on joint work with Sacha Bourgeois-Gironde. We develop a unified spatial model of turnout and voting behaviors that pertains to explain the behavior of conflicted voters, i.e., of voters that identify with two groups or parties. These are voters that have two conflicting preferences (as identities), and that, based on the aversion to betray either of their identifications, wish to satisfy both preferences. Under these conditions, we show that if there is no position that reconciles the ideological views of the two parties it is always rational for conflicted voters to abstain. Since this holds even if they could, as a group, influence the result of the election, we call it a conflicted voter's curse. In a two-candidates electoral competition, this curse implies that candidates converge to the preferred outcome of conflicted voters if and only if these voters are pivotal and the parties have shared ideological views. Otherwise, we show that convergent and divergent equilibriums are possible depending upon the degree of party polarization and if candidates care about ideology or not. These results illustrate how the behavior of some voters with multiple preferences

or identifications may influence electoral outcomes, and suggest that more research should focus upon the mixed and moderate voters that compose the political center.

Finally, Chapter 5 is based on joint work with Nobuyuki Hanaki and Benoît Tarroux, and studies multiple motivations in the data. We design an experimental study that distinguishes between different motivations that give rise to a preference for holding control in a principal-agent interaction. In particular, we refine a recent experiment by Bartling, Fehr and Herz (2014), in which they found that Swiss individuals attach an economically meaningful intrinsic value to make a decision by themselves rather than delegating it to another person. We introduce a series of treatments in order to disentangle how much of such value stems from (i) a preference for independence from others, (ii) a desire for power, or (iii) other motives such as a preference for self-reliance. In addition, we conduct a cross-cultural comparison between France and Japan to shed some light into the social determinants of such preferences. Our main findings suggest that (i) Japanese and French individuals intrinsically value decision rights beyond their instrumental benefit, that (ii) this value is greater for French than Japanese individuals, and that (iii) self-reliance is the only rationale behind the intrinsic value of decision rights in both France and Japan. We also have mild evidence that while French principals are indifferent with respect to independence and power as motives for the intrinsic value of holding control, they are negatively valued by Japanese principals. Although our experiment is not designed such that we are able to ascertain if each individual is motivated by more than one independent preference/motivation, it suggests that this might be the case for Japanese individuals that seem to intrinsically value self-reliance positively and independence and power negatively.

0.0.4 Research Methodology

In this part of the introduction I give a brief account of the epistemological view adopted in this thesis *and* the main two methods, theoretical and experimental analyses, used in the five chapters. I start with the later.

0.0.4.1 Methods

The research carried out in this thesis relies mostly upon two methods: (i) building theoretical models and (ii) conducting laboratory experiments.

Theoretical Analysis. Theoretical reasoning and modeling has a long tradition in economics. Reasoning through models has several advantages. For example, Walliser (2007) argues that a model has six functions: the *iconic* (contextualization, symbolization, and interpretation of economic phenomenon into a rigorous language), the *syllogistic* (explication, inference, and simulation of economic phenomenon), the *empirical* (confrontation and validation of theoretical ideas against empirical data), the *heuristic* (stabilization and evolution of knowledge), the *praxiological* (instrument for prediction and set of action), and the *rhetorical* one (concise expression, vulgarization, and transmission of knowledge).

At the same time, models can be highly reduced and in some occasions sparsely connected with reality. According to some schools of thought, this is an important deficiency of economic models. With the advent of big data and other empirical developments, theoretical reasoning and modeling seems to have become less prominent in recent years.[11] As defended below, I believe that theoretical models are useful to gather valuable insights about economic behavior.

Experimental Analysis. The second main method used in this thesis is the design and conduct of laboratory experimental analysis. This method has the advantage (against other empirical methods) of creating a controlled environment that is suitable to isolate and study a limited set of effects and causal relations. A laboratory

[11] Though I have not strong evidence for this claim, see e.g. Noah Smith's blog entry (available at http://noahpinionblog.blogspot.fr/2013/08/the-death-of-theory.html?m=1).

experiment "is a simple and controlled mini-world in contrast to the complex and uncontrolled maxi-world" (Maki 2005, 306). As a consequence, experimentation can bring valuable insights into how people behave and what they value, and how and why they do so.

One potential disadvantage of the experimental method is the possible low external validity of the results, i.e., their low applicability to "real world" contexts (see e.g. Guala 1999; Loewenstein 1999; Starmer 1999). For example, without knowing *why* some behavioral outcome has been obtained within an experimental setting it may be difficult to use the results beyond the context where the experiment has been ran. In the case of randomized controlled trials (RCTs), a specific type of field experiments commonly used in development settings, Deaton (2010, 448) argues that "[f]or an RCT to produce "useful knowledge" beyond its local context, it must illustrate some general tendency, some effect that is the result of mechanism that is likely to apply more broadly". The author argues that experiments should be theory-driven, and gives the example of many of the experiments in behavioral economics. Still, the step from the laboratory to other contexts may be a difficult one for other reasons. For example, it may be difficult to create "parallel" circumstances in the laboratory to the particular part of the economic system that the experiment is intended to mimic.

Taking the advantages and caveats of laboratory experimental analysis into consideration, lab experiments seem specially suited for cases in which it is difficult to isolate or identify the effects or casual relations of interest in the real world. The separation of the instrumental and intrinsic values associated with some economic or social behavior, as it is tried in Chapter 5, seems to correspond to such a case. Laboratory experiments seem specially appropriate, according to this view, to shed some light into the multiple motivations that are behind relevant economic and social behaviors and that are difficult to disentangle in other contexts.

0.0.4.2 Epistemological Statement

In recent years, a fruitful debate has surrounded the question about the epistemology of economic models (see e.g. Maki 1994; Sugden 2000; Rubinstein 2006; Gilboa, Postlewaite, Samuelson and Schmeidler 2014). An important dimension of this discussion has been the definition of what consists a good (theoretical) model. For example, Rubinstein (2006, 881) argues that a *good* theoretical model is like a *fable*, i.e., a simplified (possibly unrealistic) parallel to a situation of the real world that "identifies a number of themes and elucidates them". According to Rubinstein (2006) models are not meant to be testable and are of limited scope. They do not influence the real world through sound advice or predictive capacity, but rather by influencing "an accepted collection of ideas and conventions that influence the way people think and behave" (Rubinstein 2006, 882). Sugden (2000), on the other hand, considers that a good theoretical model is a *credible world*, i.e., a parallel reality to the real world that, given our knowledge about "the general laws governing events in the real world", could itself be accepted as real. According to Sugden (2000) "the gap between model world and real world can be filled by inductive inference".[12] What is important, according to this view, is to "recognize some significant similarity between those two worlds" (Sugden 2000, 23).

Models as Fishing Rods. My own view is somewhere in between these two. The perspective taken in this thesis is that a model (here defined as any theoretical or experimental framework) is a tool for gathering *insights*.[13] By insight I mean the result of apprehending (i) a more precise or intuitive understanding of the nature of an effect or a causal relation that may be observed in the real world, or (ii) a more precise or intuitive understanding of some economic notion or of the real world itself. Such insight might be a "general tendency" in the sense of Deaton's quote

[12] Induction is defined as any mode of reasoning which departs from specific propositions to arrive to more general ones.

[13] I believe that the view exposed here is better adapted for positive than to normative economics, and particularly suited to behavioral models (either theoretical or experimental) of the kind developed in this dissertation. See Maki (2005) for an essay in favor of seeing theoretical models as "thought" experiments and experiments as "material" models.

(see previous section), but also an understanding of some general laws through the contextualization, symbolization, and interpretation of economic phenomenon into a rigorous setup, either theoretical or experimental. For example, the experimental frameworks on voting behavior and minimal social identities have brought insights (in the form of a more precise understanding) about the nature of the potential effects of social identities on voting behavior in the real world (e.g. Schram and Sonnemans 1996; Feddersen, Gailmard and Sandroni 2009; Bassi, Morton and Williams 2011).

I believe that defining a model's *purpose* as gathering insights is suitable for positive economic analysis. According to this perspective, a model (either theoretical or experimental) that brings no insights is a useless model. For example, although being credible may enhance the capacity of a model to bring valuable insights, it seems in principle possible (although unlikely) that a credible model may be devoid of relevant insights for the real world. Similarly, a model that has a different purpose than gathering insights is, according to this perspective, a potentially useful but possibly inadequate model. For example, a fable that elucidates a given behavioral effect but that has as its main purpose to impart, in disguise, some moral point of view seems to be an inadequate model.

Many models will also be *contextual*, in the sense that causal relations in the real world are rather relative than absolute.[14] This seems to be the case for positive behavioral models, as soon as one takes a worldwide perspective; several cross-cultural experimental studies have by now documented significant differences in preferences and behavior across societies (see Chapter 5 for references). Others, such as models of personal identity, may well pertain for some kind of absolute insight. But having the contextual feature of models in mind seems to be a useful tool of interpretation.

A good model, according to this perspective, is one that brings valuable insights for a given group of people, a given context, and/or time. It may be either a fable or a credible world. In my view, either to adopt a fable or a credible world to eluci-

[14] See Guala (1999) for a similar argument within a discussion of the external validity of experimental results.

date a given subject depends upon which one is more adapted as a tool for gathering insights to that subject. Sugden (2000) makes a strong case, in my opinion, that credibility and inductive inference may favor this goal. But a (non-credible) fable may be still useful for gathering insights, for example, when these insights are intuitive understandings of the real world itself.

Finally, I believe that, much like "material" (experimental) models are tested, some premises and predictions of "thought" (theoretical) models can and should be tested. In particular, by doing so one can bring additional evidence on if its insights are valid for a specific group of people, context, and/or time. Indeed, the process of building and testing models seems to be a combined and synergistic one.

I believe that this modest perspective is adequate for economics given the tendency to judge economic (positive) models according to their empirical consistency and predictive capacity, and the frequent overconfidence with regard to the descriptive and predictive capacity of these models.[15] Under this light, the economist is the fisherman, the model is his fishing rod, and insights are just the best catch he can hope for.

0.0.5 General Notation

In the following, I define a *binary relation* $\succsim$ on any set Ω as a subset of $\Omega \times \Omega$. Following the convention in economics, I write $x \succsim y$ instead of $(x, y) \in \succsim$. Given a binary relation $\succsim$, we define its *symmetric factor* $\sim$ by $x \sim y \iff x \succsim y$ and $y \succsim x$ and its *asymmetric* factor $\succ$ by $x \succ y \iff x \succsim y$ and not $(y \succsim x)$. A binary relation $\succsim$ on Ω is said to be:

(*i*) *reflexive* if the statement $x \succsim x$ holds for every x in Ω;

(*ii*) *transitive* if $x \succsim y$ and $y \succsim z \Rightarrow x \succsim z$ for any x, y, $z \in \Omega$;

(iii) *complete* if $x \succsim y$ or $y \succsim x$ holds for every distinct x and y in Ω;

[15] See Gabaix and Laibson (2008) for an essay on seven properties of good models that include *empirical consistency* (in terms of the strength/non falsification of their predictions) and *predictive precision*. See Angner (2006) for an essay on the overconfidence of economists when acting as experts in matters of public policy.

(iv) *antisymmetric* if $x \sim y \Rightarrow x = y$, and

(v) *acyclic* if $x_1 \succ ... \succ x_k \Rightarrow \neg(x_k \succ x_1)$ for any x_1, ..., $x_k \in \Omega$.

An *ordering* on X is a reflexive, complete, and transitive binary relation on X, and a *linear ordering* on X is an antisymmetric ordering on X. Given two binary relations $\succsim_1$ and $\succsim_2$, we say that $\succsim_2$ is an extension of $\succsim_1$ (or is compatible with $\succsim_1$) if it is the case that, for any x and y in Ω such that $x \succsim_1 y$ one has also $x \succsim_2 y$. Given a binary relation $\succsim$ on a set Ω, I define its *transitive closure* $\widehat{\succsim}$ by $x \widehat{\succsim} y \iff \exists\{x_t\}_{t=0}^{l}$ for some integer $l \geq 1$ satisfying $x_t \in \Omega$ for all $t = 0, ..., l$ for which one has $x_0 = x$, $x_l = y$ and $x_j \succsim x_{j+1}$ for all $j = 0, ..., l-1$. It is well-known that the transitive closure of a binary relation $\succsim$ is the smallest (with respect to set inclusion) transitive binary relation compatible with $\succsim$.

Bibliography

Aizerman, M. A. and A. V. Malishevski (1981) General Theory of Best Variants Choice: Some Aspects. IEEE Transactions on Automatic Control 26(5): 1030–41.

Alós-Ferrer, C. and F. Strack (2014) From Dual Processes to Multiple Selves: Implications for Economic Behavior. Journal of Economic Psychology 41: 1–11.

Ambrus, A. and K. Rozen (2013) Rationalising Choice with Multi-Self Models. The Economic Journal 125: 1136–56.

Angner, E. (2006) Economists as Experts: Overconfidence in Theory and Practice. Journal of Economic Methodology 13(1): 1–24.

Apesteguia, J. and M. A. Ballester (2013) Choice by Sequential Procedures. Games and Economic Behavior 77(1): 90–99.

Apesteguia, J., M. A. Ballester, and J. Lu (2017) Single Crossing Random Utility Models. Econometrica 85(2): 661–74.

Barberà, S. and P. K. Pattanaik (1986) Falmagne and the Rationalizability of Stochastic Choices in Terms of Random Orderings. Econometrica 54(3): 707–15.

Bartling, B., E. Fehr, and H. Herz (2014) The Intrinsic Value of Decision Rights. Econometrica 82(6): 2005–39.

Bassi, A., R. B. Morton, and K. C. Williams (2011) The Effects of Identities, Incentives, and Information on Voting. The Journal of Politics 73(2): 558–71.

Becker, G., M. DeGroot, and J. Marschak (1963) Stochastic Models of Choice Behavior. Behavioral Science 8: 41–55.

Benjamin, D. J., J. J. Choi, and A. J. Strickland (2010) Social Identity and Preferences. The American Economic Review 100(4): 1913–28.

Bernheim, B. D. and A. Rangel (2004) Addiction and Cue-Triggered Decision Processes. The American Economic Review 94(5): 1558–90.

Cherepanov, V., T. Feddersen, and A. Sandroni (2013) Rationalization. Theoretical Economics 8(3): 775–800.

Deaton, A. (2010) Instruments, Randomization, and Learning about Development. Journal of Economic Literature 48: 424–55.

Dietrich, F. and C. List (2013) Where do Preferences Come From?. International Journal of Game Theory 42(3): 613–37.

——— (2016) Reason-based choice and context-dependence: An explanatory framework. Economics and Philosophy 32(2): 175–229.

Dijksterhuis, A., R. Spears, T. Postmes, D. Stapel, W. Koomen, A. van Knippenberg, and D. Scheepers (1998) Seeing one thing and doing another: Contrast effects in automatic behavior. Journal of Personality and Social Psychology 75(4): 862–71.

Eliaz, K., M. Richter, and A. Rubinstein (2011) Choosing the Two Finalists. Economic Theory 46: 211–19.

Elster, J. (1985) Introduction. In: J. Elster (ed) The Multiple Self. Cambridge University Press, Cambridge: 1–34.

Feddersen, T., S. Gailmard, and A. Sandroni (2009) Moral Bias in Large Elections: Theory and Experimental Evidence. American Political Science Review 103(2): 175–92.

Fishburn, P. (1999) Stochastic Utility. In: S. Barberà, P. Hammond, and C. Seidl (eds) Handbook of Utility Theory. Vol. 1 Principles Kluwer Academic Publishers, Dordrecht, Holland: 273–320.

Frankfurt, H. G. (1971) Freedom of the Will and the Concept of a Person. The Journal of Philosophy 68(1): 5–20.

Fudenberg, D. and D. K. Levine (2006) A Dual Self Model of Impulse Control. The American Economic Review 96: 1449–76.

——— (2012) Timing and Self-Control. Econometrica 80(1): 1–42.

Gabaix, X. and D. Laibson (2008) The Seven Properties of Good Models. In: A. Caplin and A. Schotter (eds) The Foundations of Positive and Normative Economics. Oxford University Press, New York: 292–99.

Gallagher, S. (2000) Philosophical Conceptions of the Self: Implications for Cognitive Science. Trends in Cognitive Sciences 4(1): 14–21.

Gigerenzer, G. and R. Selten (2000) Bounded Rationality: The Adaptive Toolbox. MIT Press, Cambridge, MA.

Gilboa, I., A. Postlewaite, L. Samuelson, and D. Schmeidler (2014) Economic Models as Analogies. The Economic Journal 124: 513–33.

Guala, F. (1999) The Problem of External Validity (or "Parallelism") in Experimental Economics. Social Science Information 38: 555–73.

Gul, F. and W. Pesendorfer (2001) Temptation and Self-control. Econometrica 69(6): 1403–36.

——— (2004) Self-control and the Theory of Consumption. Econometrica 72(1): 119–58.

——— (2005) The Revealed Preference Theory of Changing Tastes. The Review of Economic Studies 72(2): 429–48.

——— (2006) Random Expected Utility. Econometrica 74(1): 121–46.

Hirschman, A. O. (1977) The Passions and the Interests: Political Arguments for Capitalism Before its Triumph. Princeton University Press, New Jersey.

——— (1984) Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse. The American Economic Review: Papers and Proceedings 74(2): 89–96.

LeBoeuf, R. A., E. Shafir, and J. B. Bayuk (2010) The Conflicting Choices of Alternating Selves. Organizational Behavior and Human Decision Processes 111(1): 48–61.

LeBouef, R. A. and Z. Estes (2004) "Fortunately, I'm no Einstein": Comparison Relevance as a Determinant of Behavioral Assimilation and Contrast. Social Cognition 22(6): 607–36.

Loewenstein, G. (1999) Experimental Economics from the Vantage Point of Behavioural Economics. Economic Journal 109(453): 25–34.

Loomes, G. and R. Sugden (1995) Incorporating a Stochastic Element Into Decision Theories. European Economic Review 39: 641–48.

Maki, U. (1994) Isolation, Idealization and Thruth in Economics. In: Idealization VI: Idealization in Economics, Poznan Studies in the Philosophy of the Sciences and the Humanities. 38 Rodopi, Amsterdam: 147–68.

——— (2005) Models are Experiments, Experiments are Models. Journal of Economic Methodology 12(2): 303–15.

Manzini, P. and M. Mariotti (2007) Sequentially Rationalizable Choice. The American Economic Review 97(5): 1824–39.

——— (2012) Choice by Lexicographic Semiorders. Theoretical Economics 7: 1–23.

Marks, S. R. and S. M. MacDermid (1996) Multiple Roles and the Self: A Theory of Role Balance. Journal of Marriage and Family 58(2): 417–32.

McFadden, D. and M. K. Richter (1990) Stochastic Rationality and Revealed Stochastic Preference. In: J. S. Chipman, D. McFadden, and M. K. Richter (eds) Preferences, Uncertainty, and Optimality: Essays in Honor of Leo Hurwicz. Westview Press: Boulder, CO, 161-186., Boulder, Colorado: 163–186.

Oyserman, D., K. Elmore, and G. Smith (2012) Self, Self-Concept, and Identity. In: M. R. Leary and J. P. Tangney (eds) Handbook of Self and Identity. The Guilford Press, New York: 69–104.

Rubinstein, A. (2006) Dilemmas of an Economic Theorist. Econometrica 74(4): 865–83.

Schelling, T. (1984) Choice and Consequence: Perspectives of an Errant Economist. Harvard University Press, Cambridge, MA.

Schram, A. and J. Sonnemans (1996) Why People Vote: Experimental Evidence. Journal of Economic Psychology 17: 417–42.

Schubert, T. W. and M. Hafner (2003) Contrast from Social Stereotypes in Automatic Behavior. Journal of Experimental Social Psychology 39(6): 577–84.

Sen, A. K. (1971) Choice Functions and Revealed Preferences. Review of Economic Studies 38: 307–17.

——— (1977) Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. Philosophy and Public Affairs 6(4): 317–44.

Starmer, C. (1999) Experiments in Economics: Should We Trust the Dismal Scientists in White Coate?. Journal of Economic Methodology 6: 1–30.

Steedman, I. and U. Krause (1985) Goethe's Faust, Arrow's Possibility Theorem and the Individual Decision-taker. In: J. Elster (ed) The Multiple Self. Cambridge University Press, Cambridge: 197–231.

Sugden, R. (2000) Credible Worlds: The Status of Theoretical Models in Economics. Journal of Economic Methodology 7(1): 1–31.

Thaler, R. H. and H. M Shefrin (1981) An Economic Theory of Self Control. Journal of Political Economy 89(2): 392–406.

Tversky, A. (1969) Intransitivity of Preferences. Psychological Review 76(1): 31–48.

Tversky, A. and D. Kahneman (1991) Loss Aversion in Riskless Choice: A Reference-dependent Model. The Quarterly Journal of Economics 107(4): 1039–1061.

Walliser, B. (2007) Les Fonctions des Modèles Économiques. In: A. Leroux and P. Livet (eds) Leçons de Philosophie Économique Tome III: Science Économique et Philosophie des Sciences. Economica, Paris: 285–302.

Wheeler, S. C., K. G. DeMarree, and R. E. Petty (2007) Understanding the Role of the Self in Prime-to-Behavior Effects: The Active-Self Account. Personality and Social Psychology Review 11(3): 234–61.

Chapter 1

Mistakes or Reflexive Preferences?

Neoclassical economics uses the maximization of a stable and exogenous preference relation as the benchmark for positive and normative economics. Following the evidence that behavior depends on context and experience, several authors have designed models that treat behavior inconsistent with the maximization of a stable preference as mistaken. In this chapter, I argue that it is important to distinguish mistakes from inconsistent behavior that results from preferences (or preference change) that individuals identify with (*reflexive preferences*). I sketch two hierarchical preferences models that represent some of these ideas, and discuss how they could relate with the conflicted agent and evolving agent models in order to represent reflexive and non-reflexive preference change. Finally, I argue that collecting information on if individuals identify or not with their preferences (or preference change) may be useful for normative analysis, in particular as a refinement to welfare rankings currently used in behavioral welfare economics.

Keywords: Reflexive preferences; Mistakes; Preference change; Hierarchical preferences; Behavioral welfare economics.

1.1 Introduction

At least since Arrow (1951), it has been standard practice in neoclassical economics to assume that all tastes, values, or other preferential considerations of an individual can be summarized in a single ordering over all relevant alternatives.[1] In most economic theory and application, this ordering is taken to be exogenous and stable over time. Positive economics has viewed the maximization of this single preference as the main driving force underlying individual behavior. In normative economics, these preferences are the main ingredients for evaluating the desirability of alternative states of affairs.

The findings of psychology and behavioral economics show however that models based on an exogenous and stable preference relation are often at odds with the dependence of behavior on context and experience.[2] As a response, many authors have designed models that treat choice behavior that is inconsistent with the maximization of a stable preference as errors or mistakes.[3] These models often assume that individuals have a "true" underlying preference that they would follow would their reasoning not been distorted by a faulty psychological mechanism, or, in the absence of such an assumption, that the best for these individuals would be to follow such "consistent" preference relation.

In this chapter, I argue that it is important to distinguish between the "inconsistent" behavior that results from preferences (or preference change) that the individuals *identify with*[4] and the "inconsistent" behavior that results from preferences that the individuals do not identify with. The difference is that while the later are prefer-

[1] In Arrow (1951, 17) "[i]t is assumed that each individual in the community has a definite ordering of all conceivable social states [alternatives], in terms of their desirability to him. It is not assumed here that an individual's attitude toward different social states is determined exclusively by the commodity bundles which accrue to his lot under each. It is simply assumed that the individual orders all social states by whatever standards he deems relevant." I thank Nicolas Gravel for this reference.

[2] See Hoff and Stiglitz (2016) for a recent review and taxonomy of many of these findings and the strands of literature associated with them.

[3] See Rabin (2013) for a recent review.

[4] An individual identifies with a preference when (roughly) she evaluates or judges this preference positively, endorses it, and/or wants it to be her will. I discuss the philosophical basis of this notion in Section 1.8.2. To avoid awkward wording, I refer to individuals in the feminine.

ences that the individuals do not endorse (and that lead to choices that are judged as mistakes by the individuals themselves), the former are preferences that the individuals endorse even if they lead to behavior that is inconsistent with the maximization of a stable preference relation. I argue that this distinction may be useful for positive economics, since it may lead to better description and prediction of economic behavior. Since preferences that individuals identify with are based on their evaluation of themselves, in what follows I call them *reflexive preferences*.

I sketch two hierarchical (preferences) models that represent some of these ideas. The hierarchical *retrospective model* takes a backward-looking perspective, where preferences (choices) are judged by the individual *ex-post*. The hierarchical *evolving model* takes a forward-looking perspective, where choices are the result of reflexive or non-reflexive preferences. "Inconsistent" behavior may result from reflexive or non-reflexive preference change, i.e., preference change that an individual identifies with or preference change that an individual does not identify with respectively. I argue that behavior that is inconsistent with the maximization of a stable preference but that individuals identify with (reflexive preference change) is *not* the result of a mistaken psychological mechanism.

I then distinguish two representations of the economic agent that could be coupled with a hierarchical model to represent reflexive and non-reflexive preferences and preference change. One is based on (i) the conflict between multiple preferences and the other on (ii) the evolution of a single preference. The first refers to the cases in which an individual identifies (or not) with several motivations, roles, or goals that lead her to "alternate" between different preferences. The second refers to the cases in which an individual changes her single preference over time according to her experience and identifies (or not) with these changes.

This view is intimately linked with a *person*'s reflexive ability to form preferences over preferences, or what philosophers often call second-order desires, volitions, or

preferences.[5] I argue that data on second-order preferences may be useful for normative analysis, since it adds information on what people *identify* with, what people *value*, what they *care* about, and *who* they wish to be or become. I discuss how such information can be used as a *refinement* to welfare rankings currently used in behavioral welfare economics (e.g. Bernheim and Rangel 2007, 2009), and rejoin some of the criticisms to the use of second-order preferences in economics.

The remainder of the chapter is organized as follows. I start with a review of some of the arguments in favor of taking the maximization of a stable and context-independent preference relation as a benchmark in positive economics, notably the ones put forward by Hausman (2012) in support of sticking to a notion of preference as a *total subjective comparative evaluation* (Section 1.2). I follow with a review of the strategy of extending this approach to include mistakes (Section 1.3), and argue that it is important to distinguish between preferences that individuals identify with and preferences they don't identify with (Section 1.4). I then formalize some of these notions in Section 1.5, and discuss two conceptions of the economic agent that could be used to represent reflexive and non-reflexive preferences and preference change (Section 1.6). I continue with an appraisal of the use of preferences over preferences as a tool for welfare analysis (Section 1.7), and discuss some of the features, limitations, and extensions of this approach in Section 1.8. I conclude with a brief comment (Section 1.9).

1.2 The Rational Agent

> "[o]ne does not argue over tastes for the same reason that one does not argue over the Rocky Mountains - both are there, will be there next year, too, and are the same to all men." (Stigler and Becker 1977, 76).

[5] See Frankfurt (1971) for the philosophical basis of higher-order desires and volitions. See Jeffrey (1974) for a first treatment of second-order preferences. See e.g. Sen (1977) and George (1984) for treatments of preferences over preferences in economics.

In neoclassical economics' textbooks and most theory and application agents are assumed to have stable and exogenous preferences over all relevant alternatives. This means that whether an agent prefers x to y remains stable over time and across contexts, and that preferences are taken to be an essential but unexplained feature of the economic agent's identity. In particular, the agent never changes his or her "true" fundamental preferences over fully specified outcomes.[6] If $\mathcal{T} = \{1, ..., T\}$ denotes a sequence of periods and $\succsim$ denotes such preference, this means that if $x \succsim y$ in period t then $x \succsim y$ for every other $t' \in \mathcal{T}$, where the statement $x \succsim y$ can be read as "x is preferred or indifferent to y". Stigler and Becker (1977, 76) illustrate this view: according to the authors, economists should "treat tastes [preferences in their paper] as stable over time", and "search for differences [changes] in prices or incomes to explain any differences or changes in behavior".

The observational implications of this model are described by the traditional revealed preference axioms.[7] For instance, the Weak Axiom of Revealed Preference (WARP) states that if an alternative x is "revealed preferred" to y (i.e., x is once chosen when y is available *and* rejected), then y is *not* revealed to be "at least as good as" x (i.e., y is never chosen when x is available). This axiom is necessary and sufficient for a choice function to be rationalized by the maximization of a single preference (see e.g. Sen 1971). In terms of consumption decisions (i.e., choices over budget sets), the Generalized Axiom of Revealed Preferences (GARP, a stronger condition than WARP) is necessary and sufficient for "rational" behavior (see Afriat 1967).

In a recent book, Daniel Hausman (2012) provides an appraisal of the economists' rational agent model that has attracted a lot of attention in the literature (see e.g. Infante, Lecouteux and Sugden 2016). He is interested in describing how the concepts of *preference, value, choice, and welfare* are and ought to be used in economics, and

[6] The neoclassical approach is compatible with changes in preferences over *uncertain prospects* following an update in the agent's beliefs about the likelihood of the possible outcomes of those prospects. To be self-contained, I mostly abstract from questions related to risk and uncertainty.

[7] Examples of these axioms include the Weak and Strong Axioms of Revealed Preference and the Weak and Strong Congruence Axioms. See Sen (1971) for a seminal contribution and Varian (2006) for a recent review.

provides, in my view, an interesting reason-based conception of preference (even though, as it will become clear, I endorse an alternative way to conceptualize preference). The author argues that the concept of a single preference, as employed in neoclassical economics, is (and ought to be) a total subjective comparative evaluation (TSCE). It is *comparative* in the sense that people prefer one state of affairs to another. It is *subjective* in the sense that the comparison is made from a first-person perspective. And it is *total* in the sense that it is a comparison that takes into account everything that the economic agent considers to be relevant for choice. In the words of Baigent (1995, 92), who shares the same view as Hausman (2012) on this point, "[w]hat is being assumed is that agents who have multiple cares and concerns have resolved any conflicts into an 'all-things-considered preference'." It seems clear that preferences are comparative and subjective, and nowadays, I think most economists would agree that the concept of preference, as used in neoclassical economics, is most often an all-things considered ranking of alternatives.[8]

But according to Hausman (2012) a preference is (and ought to be) also an *evaluation*, in the sense that it is the result of a rational deliberation about what agents have most reason to do. Hausman (2012) argues that a preference is (or should be seen as) a reason-based evaluation rather than a judgment, rather than an expression of taste, or rather than a feeling, because judgments do not by themselves motivate action, tastes do not exhaust the considerations relevant to choice, and feelings - alone - do not provide *reasons* for action (see also Hausman 2013, 219). This means that an agent's choice that is not based on a rational deliberation about what she has most reason to do, such as a choice based on *intuition* (defined as an automatic impression), is not considered to reveal a preference according to Hausman's (2012) definition. As I will try to motivate in what follows, it is possible to keep the advan-

[8]Taking preferences to be *total* answers directly to the claim that a single preference is not able to incorporate an array of different motivations, cares, or concerns including moral sentiments. But assuming that an agent is able to perform a total comparison is not innocuous. For instance, people may not be able to resolve the conflict between different or opposite concerns, cares, or motivations. These may not be commensurable in the sense that trade-offs are not possible between the different rankings. If this is the case, a complete ranking of alternatives may not be achieved.

tages of seeing a preference as a total subjective comparative evaluation even if the evaluation is not necessarily based on a rational deliberation.

According to Hausman (2012, 14-20) taking preference to be a TSCE is consistent with two implicit assumptions imposed on preferences in economics (and that are relevant to my analysis).[9] First, Hausman (2012) argues that the fact that preferences take all relevant considerations for choice into account accords with the long-held idea in economics that preferences determine and motivate choices (when, as Hausman stresses, combined with beliefs). The assumption that preferences determine choices means that among the alternatives they believe to be available, the economic agent will choose one that is at the top of their preference ranking. Still, Hausman (2012, 20) argues "[t]hat choices be determined by preferences is *not* demanded by rationality". For instance, it may be rational to use a heuristic to arrive to a choice (e.g. to avoid cognitive costs imposed by reason-based deliberation). According to Hausman (2012), rationality only demands that one *should* not choose x when y is available and one is confident that all-things considered y is preferred to x. But one can argue that intuitions, *feelings* (defined as emotional reactions), or *inclinations* (defined as idiosyncratic tastes) can provide rankings of alternatives that, in given choice situations, are all that is relevant for choice. Then, remark that if one "enlarges" the concept of preference beyond reason-based deliberation (e.g. to include preferences exclusively based on inclinations or feelings) preferences would *determine* choices more generally.

Second, Hausman (2012, 16-20) defends that preferences as TSCEs are "context-independent" departing from the idea that an alternative is supposed to specify everything "relevant" to preference (given the T in TSCE). The relevant characteristics, according to Hausman (2012), are given by whatever an individual takes into consideration in a rational deliberation on what she has most reason to do (given the E in TSCE). Hausman agrees with Broome's (1991, 103) view that "[o]utcomes should

[9] See Hausman (2012, 13-20) on how the concept of preference as a TSCE also relates and partly justifies the rationality of the standard assumptions of transitivity and completeness.

be distinguished as different if and only if they differ in a way that makes it *rational* to have a preference between them" [emphasis added]. Then, whether z (or something else deemed "irrelevant" to preference) is present should not matter for the preference between x and y. It follows, according to this view, that a TSCE (and the choices it determines) is *rational* if and only if it is context-independent. Thus, according to Hausman (2012), *rational choice* is the result of context-independent evaluations that provide reasons for choice. However, remark that it does not follow from these arguments that preferences are *stable* over time (as Hausman 2012, 16 recognizes). Take x and y to be two distinguishable alternatives for which time *per se* is irrelevant for the preference between them. Even if the preference between x and y is context-independent in Hausman's sense, this preference may change over time given changes in one's rational deliberation or, if one "enlarges" the concept of preference beyond reason-based deliberation, given potential changes in one's intuitions, feelings or inclinations.

Hausman (2012) gives four arguments in favor of sticking to his notion of preference. One of these, arguably the most important (see also Lehtinen 2012), is that, according to Hausman, only the conception of preference as a TSCE allows game theory and expected utility theory to serve their predictive and explanatory roles (see Hausman 2012, 65-70). Succinctly, the rationale of this advantage in terms of game theory is that if preferences are not considered as total comparisons then the game is incorrectly specified. That is, if economists do not include all the motivating factors into the payoffs of a game then the game does not correspond to the one that is actually being played. It follows that the analysis of such incompletely specified game will provide incorrect predictions or intuitions. Remark, once more, that such advantage would hold even if preferences are not seen as reason-based evaluations in the sense of Hausman (2012). What is needed is that preferences are total comparisons.

The other three arguments given by Hausman (2012, 64-5) are that a TSCE (i) matches economic practice, that (ii) it conforms roughly with the everyday usage of the word preference which helps avoid misunderstandings, and that (iii) it allows to

pose questions concerning what preferences depend on. As noted by Lehtinen (2012), these arguments refer to pragmatic advantages, and for that reason carry low weight in deciding which notion of preference to adopt as a benchmark in economics. In my view, the third reason, although pragmatic in nature, is more relevant than the two former in that respect. In particular, it points out how sticking to Hausman's notion of preference as a benchmark for economic analysis may be useful to separate different notions and aims of research. In Hausman's (2012, 65) view, "[b]y treating preferences as total rankings, economists can separate the use of the word 'preference' from substantive views about what preferences depend on". As it becomes clear along the book, Hausman (2012) believes that economists should focus on how preferences are formed, and thinks that adopting a notion of preference as a TSCE is useful since, unlike some other notions of preference - such as preferences as exclusive expressions of tastes -, "it does not settle *a priori* what influences preferences" (p. 65). It suggests instead that preferences can be determined by several motivations. As before, this advantage seems to be shared by any definition of preference that is a total subjective comparison, but not necessarily a reason-based one.

A useful or misleading benchmark?

Hausman's (2012) arguments suggest that models based on the maximization of a single preference, when preference is seen as a stable TSCE, may be useful in terms of *parsimony*, *generality*, and *tractability*. In particular, such models allow economists to efficiently use the wide range of tools that they have developed so far, as it is the case of game theory. According to Hausman (2012, e.g. 73), this representation of human behavior and rationality is a useful benchmark when the main interest is the determination of action by the interplay of beliefs, constraints, and preferences. And, as it is the case with consumer choice theory, it seems that this is sometimes the case at least as a first approximation to the agents' choice behavior.

Still, seeing preferences as total comparative evaluations is a strong idealization. Besides excluding preferences that are not based on rational deliberation, Hausman's

conception of preference is not compatible with several sources of preference change when preferences are taken to be stable.[10] In particular, and as developed below, a stable TSCE is not compatible with changes in preferences due to changes in values or other experiences of the agent. This is problematic, among other things, because the theory is silent with respect to the effect of changes in market rules or changes in other economic institutions on preferences *per se*.

The findings of psychology and behavioral economics also suggest that neither the choice behavior nor the decision making of most individuals accord with this theory. The accumulating evidence on the context-dependency of behavior and limits to rationality bring several doubts concerning the predictive capacity and normative value of the rational agent model. Next, I turn to one of the most prominent answers among behavioral economists that tries to address these issues but that ends up "replacing" the rational agent by an *inner or outer rational agent*[11].

1.3 The Inner and the Outer Rational Agents

Behavioral economics is by now a field interested in very different determinants of behavior, but the first and leading strand of literature has focused in inconsistent choices that result from intuitive or seemingly faulty psychological mechanisms (see Hoff and Stiglitz 2016). There is a long list of studies that show the effects of frames, anchoring, inattention, and other bias on decision making. In order to accommodate these findings, many authors have proposed incremental improvements to the standard rational choice model. In a review of many of these proposals, Rabin (2013, 528) argues that such improvements incorporate greater "realism while attempting to maintain the breadth of application, the precision of predictions, and the insights of neoclassical theory".

[10] See e.g. Livet (2006) and Dietrich and List (2013, 2016) for discussions of this limitation and attempts to build theories of preference formation *and* preference change.

[11] I borrow the term "inner rational agent" from Infante et al. (2016). See their essay for a critical analysis of the approach discussed in the next section when related to welfare economics.

According to this view the maximization of a stable preference is seen as a useful benchmark or first approximation, from which behavioral theories are supposed to be judged against, for instance, with respect to their additional explanatory power. At the same time, many authors treat departures from the standard assumptions about rational choice as *mistakes* (e.g. Akerlof 1991; O'Donoghue and Rabin 2003; Bernheim and Rangel 2004). According to Rabin (2013, 529), "[w]e can capture many errors in terms of systematic mistakes in the proximate value function people maximize (quasi-maximization models)".

An example is provided by the literature on *preference reversals* in intertemporal choice between a smaller short-term reward and a larger long-term reward, that according to Rabin (2013, 534) is the most successful of the incremental improvements to neoclassical economics.[12] Many authors interpret these preferences as *present-biased*, and represent them in a two-parameter model that modifies exponential discounting (see e.g. Akerlof 1991; Laibson 1994, 1997; O'Donoghue and Rabin 1999, 2001, 2003). Let u_t be the instantaneous utility an individual derives from an activity in period $t \in \mathcal{T} = \{1, ..., T\}$. Then, these models (numerically) represent the individual's intertemporal preferences at period t with the following utility function:

$$\text{For all } t \in \mathcal{T} = \{1, ..., T\},$$
$$U^t(u_t, u_{t+1}, ..., u_T) \equiv \delta^t u_t + \beta \sum_{\tau=t+1}^{T} \delta^\tau u_\tau \tag{1.1}$$

where $\beta > 0$ and $\delta \leq 1$. Remark that δ represents "time-consistent" discounting, and that if $\beta = 1$ these preferences represent standard time-consistent exponential discounting. Instead, if $\beta < 1$ these preferences are interpreted as "time-inconsistent" preferences for instantaneous utility (i.e., preferences that are present biased in the sense that the individual gives more relative weight to period τ in period τ than she

[12] In the typical example or experiment, agents choose between a smaller reward at period 2 and a larger reward at period 3. If the choice is made at period 2, then the smaller-earlier reward is chosen. If instead the choice is made prior to period 2, then the larger-later reward is chosen.

does in any period prior to period τ). This present bias (and the preference reversal it entails) is often interpreted as a defect. For instance, O'Donoghue and Rabin (2003, 187) treat "this preference for immediate gratification as an error", and Rabin (2013, 538) regards the present bias to be a "quasi-maximization error".

Accordingly, some authors, such as Akerlof (1991), interpret an unobserved long-run intertemporal preferences (with $\beta = 1$, i.e., with the present bias "removed") as the "true" preferences of the individual. However, other authors such as O'Donoghue and Rabin (1999) do not assume that individuals have such "true" underlying preferences. Instead, they take the long-run intertemporal preferences with $\beta = 1$ to be "fictitious" and interpret it to represent the preferences that individuals *should* have would they have not been biased. O'Donoghue and Rabin (1999, 112-3) justify this assumption arguing that "[s]ince present-biased preferences are often meant to capture self-control problems, where people pursue immediate gratification on a day-to-day basis, we feel the natural perspective [for welfare analysis] in most situations is the 'long-run perspective'." What is assumed is that the *best* for individuals would have been to follow the reasoning of a rational agent, that they have not followed because they are psychologically bias or *naive*. Then, these authors associate normative authority to a given and unobserved preference that is time- and context-independent and implicitly assume that any deviation from this mode of reasoning is a mistake.

Since committing a mistake is, by definition, making an action that departs from something that is true, proper, or right, models such as O'Donoghue and Rabin (1999) implicitly assume that the true, proper, or right thing to do is to maximize an unobserved and stable preference relation. This benchmark is taken as the right thing to do even if it contradicts the preferences revealed by the individual. Moreover, contrary to Akerlof (1991), O'Donoghue and Rabin (1999) do not associate the unobserved rational preference with an underlying "true" preference of the individual. Then, models in this vein can be seen as taking an *outer rational agent* (i.e., a fully rational agent that is not part of the individual and may disagree with the individ-

ual's preferences), as the guide for *sophisticated* behavior and the source of normative authority.

Consider now another interpretation of the problem of preference reversals, that represents the intrapersonal conflict between present and future preferences with "dual-self" or "dual-mode" models (e.g. Thaler and Shefrin 1981; Bernheim and Rangel 2004; Fudenberg and Levine 2006, 2012). These models represent intertemporal decisions as intrapersonal interactions between two selves or two modes: one impulsive and myopic and the other patient and far-sighted. For instance, Fudenberg and Levine (2006) treat the two selves as two *rational* players in an intrapersonal game: a "planner" (concerned with lifetime consumption) and a "doer" (that exists only for one period and is only interested in the consumption of that period). Bernheim and Rangel (2004) build a similar model to study addictive behavior, in which according to the exposure to "environmental cues", an individual alternates between a "hot mode" in which she always takes an addictive behavior "irrespective of underlying preferences", and a "cold mode" in which she "considers all alternatives and contemplates all consequences" and selects her most preferred alternative. In this sense, Bernheim and Rangel (2004) assume that each individual has a stable and single preference relation, and that choices taken under the hot mode are "mistakes" that may differ from the choices that would be determined by the individual's preferences. Instead, Fudenberg and Levine (2006) assume that both players are rational maximizers who have the same short-run preferences, and interpret the individuals' seemingly mistakes as cases of high self-control costs. But in both cases, the long-run perspective (the planner or the cold mode) is taken as the source of normative authority.

These models are related to the two selves model recently popularized by Daniel Kahneman (2011). According to the author, human psychology can be divided into two "systems" or modes of thought: one fast, effortless, and automatic (System 1), and another slow, effortful, and controlled (System 2). These two systems correspond "roughly to intuition and reasoning", and while System 1 generates involun-

tary impressions of the objects of perceptions and thought, System 2 is involved in all intentional judgments based on impressions or deliberate reasoning (see also Kahneman 2003). System 2 is also thought to monitor the activities of System 1, and the preferences of the former are not necessarily consistent with the preferences of the later. In this light, the "doer" in Fudenberg and Levine (2006) and the "hot mode" in Bernheim and Rangel (2004) are separate systems responsible for intuitive choices, while the respective "planner" and "cold mode" are separate systems responsible for reason-based actions.

As argued by Infante et al. (2016), these models suggest that individuals are endowed with an *inner rational agent*, i.e., an independent agent that is able to make stable and context-independent decisions *exclusively* based on its own evaluations of alternatives.[13] In this view, "human psychology is represented as a set of forces which affects behaviour by *interfering with* rational choice", that is itself "represented by the error-free reasoning of the inner agent" (Infante et al. 2016, 14-5). Decisions based on a psychological bias or a System 1 (hereafter intuition) are in this perspective deemed not to reveal an authentic and/or normatively relevant preference, and the agent is often assumed to have a stable and context-independent preference based on a slow, deliberated, and controlled System 2 (hereafter reasoning).

A useful or misleading benchmark?

I have distinguished two stances. One does not assume that agents are endowed with an underlying system capable of rational decision making, while the other assumes that such an inner rational agent exists. At the same time, we have seen two common trends: (i) to treat choices that result from psychological bias or intuition as mistakes (even though some authors try to avoid this assumption as e.g. Fudenberg and Levine 2006), and (ii) to assume that choices that result from psychological bias or intuition do *not* reveal authentic and/or normatively relevant preferences. The behavior of an

[13]These are the models, as pointed by Infante et al. (2016), that are the closest to explain this assumption. See their essay for a critic of other models that implicitly assume the existence of an inner rational agent without specifying an underlying model of rationality.

inner or outer rational agent is taken as the source of normative authority, and the traditional rational agent model is used as a reference point or approximation from which extensions and variations are found and modeled.

Building models that are incremental improvements of the rational agent model is often a useful strategy in economics. Besides favoring parsimony, tractability, and generality, it allows to observationally distinguish between choices that result from a stable preference and choices that result from other factors. Models in this vein may be particularly useful in cases of choices that are uncontroversially determined by other factors than what can reasonably be associated with an individual's preferences. Bernheim and Rangel (2004, 1561-2) give the example of American visitors to the United Kingdom who suffer injuries and fatalities because they only look to the left before crossing the street, even though they know that traffic approaches from the right. It seems clear that one "cannot reasonably attribute this to the pleasure of looking left or to masochistic preferences". The behavior, in this case, can be uncontroversially considered as resulting from a mistake.

However, in the remainder of the chapter I wish to argue that interpreting any deviation of "rational" behavior as mistakes is often an overly "mechanical" recipe for both positive and normative economics. Namely, I will argue that recognizing that not all these decisions are the result of mistakes, and that individuals themselves are, *a priori*, the best (or at least the first) judges about the nature of a decision, may help economists to design better explanations and predictions of behavior as well as better welfare criteria.

1.4 Mistakes or Reflexive Decisions?

> "May I urge that changes in values do occur from time to time in the lives of individuals, of generations, and from one generation to another, and that those changes and their effects on behavior are worth exploring - that, in brief, de valoribus *est* disputandum?" (Hirschman 1984, 90)

The previous two sections raise important questions. Hausman's (2012) view in favor of a conception of preference as the result of rational deliberation and the models that treat either psychological bias or intuition as mistaken raise the following questions: (i) can and ought a choice based on intuition (be interpreted to) reveal a preference, and (ii) can and ought a choice based on a psychological bias or intuition *not* always be (interpreted as) a mistake. As for the first, recall that according to Hausman's (2012) definition of preference choices based on intuition cannot reveal a preference because such choices are not evaluations based on a *rational deliberation* about what the individual has most reason to do. However, intuitions, like feelings or inclinations, seem to be important components of individuals' evaluations of alternatives that, in some occasions, are part of a person's rational deliberation about her reasons to choose. Then, it seems strange to disregard these motivations - intuitions, feelings, or intuitions - as determinants of preferences in the occasions that they determine choices even though those choices are not the result of rational deliberation (specially in a positive theory of behavior). As argued above, allowing preference rankings to be also determined exclusively by intuitions, feelings, or inclinations is more consistent with the assumption of choice determination and would be consistent with context-independent but unstable preferences.

Moreover, upon rational deliberation I may evaluate two alternatives to be equally worthy in terms of all reasons besides an intuition, feeling, or inclination in favor of one of the two alternatives. Then, I may *prefer* x to y - according to Hausman's (2012) definition - based on a first automatic impression (intuition), feeling, or inclination. Although a choice based on this preference is not an example of a "fast" choice exclusively based on intuition (since it is made after rational deliberation), it illustrates how it is possible to reveal a preference - in Hausman's (2012) sense - for one alternative over another because of a first automatic impression.[14]

[14] Infante et al. (2016) build a related but different argument based on the possibility that, all-things considered, two alternatives may be incomparable. The authors argue that in that case it is rational to choose based on an inclination, and that for that reason the individual's choices may be context-dependent. One question that emerges is if this inclination can be separated, as Infante et al. (2016) seem to assume, from the all-things considered rational deliberation.

One could argue that the choice just described reflects one's natural tendencies or inclinations, instead of the expression of an "authentic" preference between the two alternatives. A similar argument could be made for choices exclusively based on intuition, feelings, or inclinations. Although I favor calling such tendencies and inclinations expressions of preference, I think such argument points towards the relevance of the following distinction: between the preferences/inclinations that a person identifies with and the ones that a person does not identify with. If, say, upon reflection, I identify with the choice of x over y that I made based on intuition, this choice, even if not based upon a process of slow rational deliberation, seems to have revealed a preference for x over y that I (and not an observer) deem authentic. I call such "self-authenticated" preference a *reflexive preference*.

According to this view, choices can reveal preferences when based on either reasoning or intuition, but it is important to distinguish between preferences with which people identify and preferences with which people do not identify (reflexive and non-reflexive preferences respectively).[15] This distinction can be conceptualized through people's preferences over preferences (or choices), also known as hierarchical preferences, meta-preferences, or second-order preferences. For example, "I would prefer not to prefer to smoke" (to non-smoking) is a second-order preference. And since preferences in general determine choices in economics, it is often possible to describe such preferences in relation to observed choices such as "I would prefer myself not to smoke". In the next section I "decompose" second-order preferences into two independent rankings, but before that I wish to discuss some of their features and their relation to the notions of mistake and preference change.

Three features of second-order preferences are worth mentioning. The first is that second-order preferences correspond to an important part of people's *values*[16] (see also Hirschman 1984). This is the view of Lewis (1989, 115), who argues that desiring to desire (a second-order desire) is *valuing*:

[15] Examples of choices that, according to this view, do not reveal preferences are choices based on manipulation. I discuss issues related with adaptation and false beliefs in Section 1.8.3.

[16] Values are here defined as *something* (in this case preferences) intrinsically valuable or desirable.

> "The thoughtful addict may desire his euphoric daze, but not value it. Even apart from all the costs and risks, he may hate himself for desiring something he values not at all. It is a desire he wants very much to be rid of. He desires his high, but he does not desire to desire it, and in fact he desires not to desire it. He does not desire an unaltered, mundane state of consciousness, but he does desire to desire it. We conclude that he does not value what he desires, but rather he values what he desires to desire."

The second feature is that second-order preferences are not only relevant for cases of (lack of) self-control. In many applications, the conflict between first- and second-order preferences is indeed related with addictive or impulsive behavior that leads to a lack of self-control (e.g. smoking, drugs use, betrayal). Second-order preferences are a way to rationalize and predict such behavior that in general escapes the traditional rational choice model. However, second-order preferences are also relevant to inform cases related with the values, goals, and aspirations of individuals that are not related with self-control. For example, I may want myself to prefer to do more non-paid voluntary work, but have a first-order preference for more paid work given budget constraints.

The third feature is that second-order preferences are not *direct* determinants of choices, i.e., they do not necessarily imply action.[17] For instance, many heroin addicts, even if they do not identify with their preference for heroin and would like to quit using drugs, will often fail to do it. Even in many non-addictive behaviors of our day-to-day life, we often behave in ways that we do not identify with and/or that do not accord with our values. Second-order preferences are likely to be taken into consideration in an all-things considered rational deliberation, but they do not directly imply that preferences or action will be aligned with them.

[17] According to Lewis (1989), you are disposed to follow your second-order preferences/values if you were put under hypothetical "ideal conditions" to follow them. This means that, according to the author, second-order preferences *directly* determine choices only under conditions that are not usually met in real life. See also Frankfurt (1988).

I would like now to argue that a choice resulting from a psychological bias or intuition should not always be interpreted as a mistake. As remarked before, committing a mistake is, by definition, making an action that departs from something that is true, proper, or right. A sensible, if not natural, reference of truthfulness for a person is *who* this person is and what she values. According to this view, a person's choice is a mistake if the choice does not correspond to *who* she is and/or what she values. In most cases, the best judge of who one is the individual herself. The individual is also the most likely to know her values.

My suggestion is that a mistake is an action that I judge as mistaken.[18] It follows that a choice based on a psychological bias or intuition *may not* be a mistake. For instance, I may recognize that I acted based on intuition when I bought that delicious ice-cream that ruined my diet. However, in retrospect, I may not consider it as a mistake: the diet was not reflecting, according to my own judgment, who I am, want to be or become.

Similar considerations apply to the example of preference reversals in intertemporal choice that we have seen in the previous section. It may be the case that a person does not judge her present-bias as a mistake, but instead see it positively and identifies with it. For instance, it may be the case that the person wants to live her life at the fullest and values (has a second-order preference for) higher immediate gratification against lower future utility. This example illustrates that it is different to assume *a-priori* that a present bias is a defect, then to assess if this is indeed the case according to the person's own judgments about her preferences. In this sense, second-order preferences tell us, from the individual's own perspective, if she endorses or not the present bias.

To sum up, I have argued that a choice *exclusively* based on intuition can reveal an authentic preference, and that not all choices based on a psychological bias or intuition should be treated as mistakes. I have argued that what is important is to

[18]This *self-authenticated* definition contrasts with an *objective* definition of mistakes that may include choices based on adaptation or false beliefs as seen from the point of view of the observer. I discuss these issues in Section 1.8.3.

distinguish between "self-authenticated" preferences and preferences that individuals do not identify with. It results that choices that violate the traditional axioms of revealed preference and that are often treated as mistakes may be instead the result of preference change that individuals identify with. I call such changes *reflexive preference changes*.

Reflexive preference changes can be conceptualized as preference changes that result from the resolution of a conflict between a first-order preference and a second-order preference (see also Hirschman 1984). For instance, I may have a first-order preference for eating meat but form a second-order preference that values it negatively because of ethical, environmental, or other reasons and become a vegetarian such that I align my preferences with my values.

There are several reasons why it is important to consider reflexive preferences and reflexive preference change in economics. First, recognizing that changes in preferences may be the result of reflective decisions may lead to better description and prediction of economic behavior (see also Hirschman 1984, 90). For example, distinguishing between preference change due to changes in values and preference change due to changes in tastes is important since values and tastes are not, *a priori*, susceptible to be affected similarly by changes in market rules or other economic institutions. Similarly, distinguishing choices that individuals identify with and self-authenticated mistakes may help the correct interpretation of data from the "real world" or from laboratory experiments. Second, reflexive preferences and reflexive preference change can describe and predict phenomena of interest to economics that are not captured by the traditional rational choice model (such as lack of self-control). Finally, these notions are relevant for welfare inference and policy analysis. For example, knowing if "inconsistent" choices are the result of self-authenticated mistakes or choices that the individuals identify with may help to refine behavioral welfare rankings proposed in the literature (see Section 1.7).

As argued by Hirschman (1984, 90), the possibility of reflexive preference change brings an important argument against the view, defended by Stigler and Becker

(1977), that all economic behavior change should be explained and understood through changes in prices and incomes. In fact, Stigler and Becker (1977) argued that preference change is of little interest since it often results from "capricious" changes in tastes.[19] By contrast, reflexive preference change suggests a non-capricious way by which change in preferences may occur and that, according to the view defended here, economists should not overlook.

1.5 Hierarchical Models

In what follows I sketch two *hierarchical* models. The hierarchical *retrospective model* takes a backward-looking perspective, where choices are judged by the individual *ex-post*. This model is, among other things, adapted for welfare analysis. When presenting it, I introduce some formal definitions that will be used later in the refinement of a behavioral welfare ranking (Section 1.7.2). In particular, I "decompose" second-order preferences into two independent rankings: identification *and* valuation preferences. The hierarchical *evolving model* takes a forward-looking perspective, where choices are the result of the conflict (or resolution) between first- and second-order preferences. I informally explore how this model could explain preference and choice reversals. In the next section, I discuss two models of the economic agent that could be used in combination with a hierarchical evolving model to formalize these notions in the future. Before proceeding, I introduce some general definitions and the framework of choice with time that will be used[20]

1.5.1 Framework

Let X be a universe of alternatives that are of interest to the economic agent. Different alternatives, denoted by x, y, etc., can be standard objects such as consumption

[19] Stigler and Becker (1977, 89) also argued that changing preferences provide "endless degrees of freedom". See Fehr and Hoff (2011, 398-400) for why this is not a substantive argument with today's knowledge about preference explanations.

[20] I adopt a framework of choice with time that is first developed in a joint work with Nicolas Gravel and presented later in Chapter 3 of this dissertation.

bundles, but alternatives can also include non-standard features as long as they are complete and mutually exclusive descriptions of the world. For example, an alternative may include the description of possible actions (instead of standard objects) when combined with the remaining description of the world. Let $\mathcal{P}(X)$ be the set of all non-empty subsets of X, and $\mathcal{F} \subseteq \mathcal{P}(X)$ be a collection of subsets of X. Each of these subsets is interpreted as a *choice problem* (using Arrow's 1959 terminology), sometimes called a choice situation.

In what follows, I will often refer to two-element subsets of X (hereafter binary choices). Binary choices are related to many of the examples I have provided, and are intimately linked to the concept of rational choice.[21] They also illustrate the possibility of choosing between action and inaction (e.g. between "smoking" and "not smoking" when the two alternatives differ only in this respect).

In the following, I adopt the framework that is developed in Chapter 3. Let $\mathcal{T} = \{1, ..., T\}$ denote a discrete time horizon, and $K : \mathcal{T} \longrightarrow \mathcal{F}$ a *chronology of choices* that assigns to every choice period $t \in \mathcal{T}$ a unique non-empty set $A(t)$, interpreted as the choice problem taking place at *period t*. Note that K may be any sequence of choice problems, and include the same non-empty set at two distinct choice periods. As in Chapter 3, a *chronological choice function* C is a mapping that assigns to every pair $(t, A(t))$ of a chronology K a unique element $C(t, A(t)) \in A(t)$. $C(t, A(t))$ is interpreted as the chosen alternative at $(t, A(t))$.

As in the standard theory of choice, it is possible to define what it means for an agent to be *consistent* in her choices. One definition of consistent choice, denoted $x \succsim^C y$, can be stated as follows:

Definition 1. *Alternative x is said to be **consistently** chosen over alternative y (or $x \succsim^C y$) if and only if $y \neq C(t, A(t))$ for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$ for some $A(t) \in X$ such that $x, y \in A(t)$.*

[21] See Bossert, Sprumont and Suzumura (2005, 2006) for recent work on the rationalizability of choice functions on the domain that includes all singletons and all two-element subsets of X.

In words, one alternative is consistently chosen over another if the former alternative has been chosen at least once when the latter was present and the latter has never been chosen when the former was present.

1.5.2 Hierarchical Retrospective Model

The *retrospective model* is defined over observed choices. I interpret choices to reveal the agent's ***actual preferences***[22]. Remark that actual preferences may not be the result of rational deliberation as in Hausman's (2012) sense, i.e., they may be determined by desires based on inclinations, feelings, or intuitions. In this model the reflexive attitudes take place in retrospect at period T (the end of the time horizon). This seems to be a reasonable assumption if we are interested, for example, in determining which preferences/choices to take into account in terms of individual well-being *ex-post* (see Section 1.7). It is less appealing, for example, if one is interested in explaining preference change or for the prediction of behavior from one period to another (see Section 1.5.3).

For all $(t, A(t))$, define $\succsim_t^I$ on any $A(t)$. For any $x, y \in A(t)$, $x \succ_t^I y$ if and only if the agent *identifies with the choice of* x *from* $A(t)$ *more than with the choice of* y *from* $A(t)$.[23] Judgments of the type $\succsim_t^I$ are interpreted as the agent's ***identification preferences*** at period T. They can be revealed, for example, through stated second-order attitudes such as "I would myself want to have chosen x from $A(t)$ more than have chosen y from $A(t)$". The preference $\succsim_t^I$ is not necessarily complete and transitive, but is assumed to be reflexive (as a binary relation) and acyclic. I now define what I mean by a reflexive choice:

[22] I borrow this term from Harsanyi (1997). See Section 1.8.3 below for the contrast with *informed preferences*.

[23] Remark that these preferences, though formally defined as first-order, are interpreted to be of second-order because they are over alternatives contained in past choice problems where it is assumed that choices reveal actual preferences. In this sense, $x \succ_t^I y$ can be read as if $(x \succ_t y) \succ_t^I (y \succ_t x)$. See Watson (1975, 219) for an alternative view to Frankfurt (1971) based on the perspective that some second-order attitudes can be instead seen as first-order.

Definition 2. *A choice $C(t, A(t))$ is said to be **reflexive** if and only if $C(t, A(t)) \succsim_t^I x$ for some distinct $x \in A(t)$ and $y \succ_t^I C(t, A(t))$ for no distinct $y \in A(t)$.*

In words, a choice is said to be reflexive if in retrospect the agent weakly identifies with the chosen alternative with respect to at least one other feasible alternative and for no other feasible alternative it is the case that she identifies more with that alternative than with the chosen one. In the case of a binary choice, a choice is said to be reflexive if in retrospect the agent weakly identifies with the chosen alternative with respect to the other feasible alternative. Then, for any binary choice, an agent may either identify with the chosen alternative more than with the other alternative, be indifferent in terms of identification between the two alternatives, do not identify with the chosen alternative in the sense that she identifies more with the other possible alternative, or have no identification preference between the two alternatives.

Note that one could certainly weaken or strengthen the definition of a reflexive choice. For example, one could require that the chosen alternative is strictly preferred in terms of identification to at least one alternative and/or exclude the possibility of indifference in terms of identification. Note also that I abstract from some important issues with this formulation. For example, it is possible that an agent identifies more with a chosen alternative than with another feasible alternative but not *enough* for her to identify with the choice itself. It is also possible that an agent identifies with a choice itself, though there is no feasible alternative for which she identifies less than the chosen one. But for the current purposes, I stick with this definition.

Clearly, a choice may not be reflexive in this sense. For example, a choice (or the actual preference behind it) may not agree with the identification preference over it. Whenever an agent makes a choice that is against her identification preferences, then this choice seems to be judged negatively *by the agent herself*. Then:

Definition 3. *A choice $C(t, A(t))$ is said to be a (self-authenticated) **mistake** if and only if there exists a distinct alternative $x \in A(t)$ such that $x \succ_t^I C(t, A(t))$.*

In other words, a choice is said to be a mistake if there exists at least one feasible alternative that the agent identifies more than with the chosen one. For the case of binary choices, this is the converse of a reflexive choice. In case of subsets with more than two alternatives, it is possible to have a choice that is neither reflexive nor a mistake: it suffices that $C(t, A(t)) \succsim_t^I x$ for no distinct $x \in A(t)$ and $y \succ_t^I C(t, A(t))$ for no distinct $y \in A(t)$.

I now relate Definition 1 with Definition 2. If between periods 1 and T the agent makes a choice over two alternatives based exclusively on reflexive preferences, then one can say that she identifies with *all* her choices over these two alternatives. This lead us to the following definition:

Definition 4. *Alternative x is said to be* ***reflexive-consistently chosen*** *over alternative y if and only if $x \succsim^C y$ and for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$ it is the case that $x \succsim_t^I w$ for some distinct $w \in A(t)$ and $z \succ_t^I x$ for no distinct $z \in A(t)$.*

In words, an alternative x is said to be consistently chosen in a reflexive fashion over an alternative y if x is consistently chosen over y and whenever x is chosen and y is present the choice is reflexive. In the case of binary choices, this definition is independent with respect to other alternatives when judging if the choices between two alternatives have been reflexively consistent. Otherwise it is not. An alternative definition could impose a certain independence with respect to other alternatives for any choice problem.

Besides identification preferences, the agent is assumed to have reflexive and acyclic (possibly incomplete and not necessarily transitive) preferences over her different choices. This means that the agent is able to make statements of the sort $C(t, A(t)) \succsim^V C(t, A(t'))$, where $\succsim^V$ is defined over $\cup_{t \in \mathcal{T}} C(t, A(t)) \subseteq X$. These statements are interpreted as her ***valuation preferences***, and reflect her evaluative judgments about the relative importance of different choices (or what the agent cares

about).[24] These preferences incorporate a sense of valuation that is often absent in most utility views of agency and well-being. As argued by Sen (1987, 19-20), "valuation is a *reflexive* activity in a way that 'being happy' or 'desiring' need not be". As it was the case with identification preferences, valuation preferences are taken to be the result of the agent's reflexive activity at period T. In some examples and one of the applications discussed below these preferences will not be used (Section 1.7.2). But in general, if they exist, they seem to provide valuable information on what people take to be of value or care about.

The distinction of the two types of preferences seems to be descriptively meaningful. In particular, they seem to reflect preferential (reflexive) judgments of different natures: the former concerning the identification with the alternatives that could have been chosen in each choice problem, and the latter concerning a value ranking over different choices or actual preferences. Moreover, identification preferences can relate an actual choice (e.g. "smoking" has been actually chosen over "not smoking") to a hypothetical choice ("not smoking" being chosen over "smoking", even though this has never been observed as a choice); on the contrary, valuation preferences are defined over the several actual choices that have been observed.

Note that these rankings differ from the traditional view of individual meta-preferences adopted in economics.[25] In general, meta-preferences are assumed to be an unique ordering over (hypothetical) multiple first-order preferences over the universe of alternatives. In this chapter, I interpreted two independent rankings as reflecting the agent's second-order retrospective preferences over her past preferences or choices. These rankings also differ from other meta-rankings based on morality, ideology, or political priorities, in that identification and valuation preferences are based on *individual* evaluations instead of an observer's point of view.

[24] See Decancq, Fleurbaey and Schokkaert (2015) for a similar definition over functionings instead of choices/preferences.

[25] See Sen (1977) for a brief discussion of different interpretations of meta-rankings. Although Sen (1977) focus on an observer's "moral" ranking, he notices that a meta-ranking can "be ordered also on grounds other than a particular system of morality", such as the "preferences one would have preferred to have" (1977, 338-9).

1.5.3 Hierarchical Evolving Model

Contrary to the retrospective model, in the evolving model the reflexive attitudes take place at every $t - \epsilon$ with $\epsilon \in]0, 1[$. In other words, the agent is assumed to judge the preferences over her potential choices before every period. This formulation could capture some of the essential features of *second-order volitions* (i.e., second-order desires that first-order desires effectively motivate or move you to action) that are at the heart of the philosophical thought about second-order attitudes, and that are absent from a hierarchical retrospective model. Bratman (2003, 224) summarizes the features that are common to many (hierarchical) models that consider second-order volitions in the tradition of Frankfurt (1971):

> "First, it will involve a second-order attitude that is about that desire. Second, this second-order attitude will itself be a conative attitude, in the broad, generic sense of a motivating attitude. Third, this second-order conative attitude will concern certain kinds of further functioning, from now on, of the first-order desire. The content of this second-order attitude will be in this sense forward-looking. Fourth, this forward-looking second-order conative attitude will include in its own functioning the guidance, from now on, of the functioning of the first-order desire. In short: the theory will appeal to a higher-order attitude that is conative, forward-looking in its content, and guiding in its function.
>
> There is also a fifth feature that such theories try to capture. The higher-order, forward-looking, and guiding conative attitude is to constitute - at least in part, and given relevant background conditions - a commitment on the part of the agent concerning the role of the target desire in her own agency: the agent is appropriately settled on this."

In what follows, I sketch analogous definitions to the ones of the retrospective model. These could be used for the *ex-post* evaluation of choices, but also to predict future behavior. For all $(t, A(t))$, let $\succsim_t^{I_{t-\epsilon}}$ be defined on any $A(t)$. For any $x, y \in A(t)$,

$x \succ_t^{I_{t-\epsilon}} y$ if and only if the agent *identifies with the choice of x from $A(t)$ more than with the choice of y from $A(t)$ at period $t-\epsilon$*. Judgments of the type $\succsim_t^{I_{t-\epsilon}}$ are interpreted as the agent's **identification volitions** at period $t-\epsilon$, i.e., conative, forward-looking, guiding and committed attitudes in favor of choosing one alternative over others at period t. The preference $\succsim_t^{I_{t-\epsilon}}$ is again assumed to be reflexive and acyclic, but not necessarily complete or transitive. They can be recovered, for example, through stated second-order attitudes at period $t-\epsilon$ such as "I would myself want to choose x from $A(t)$ more than choose y from $A(t)$". Then:

Definition 5. *A choice $C(t, A(t))$ is said to be ϵ-**reflexive** if and only if $C(t, A(t)) \succsim_t^{I_{t-\epsilon}} x$ for some distinct $x \in A(t)$ and $y \succ_t^{I_{t-\epsilon}} C(t, A(t))$ for no distinct $y \in A(t)$.*

Definition 6. *A choice $C(t, A(t))$ is said to be a (self-authenticated) ϵ-**mistake** if and only if there exists a distinct alternative $x \in A(t)$ such that $x \succ_t^{I_{t-\epsilon}} C(t, A(t))$.*

Definition 7. *Alternative x is said to be ϵ-**reflexive-consistently chosen** over alternative y if and only if $x \succsim^C y$ and for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$ it is the case that $x \succsim_t^{I_{t-\epsilon}} w$ for some distinct $w \in A(t)$ and $z \succ_t^{I_{t-\epsilon}} x$ for no distinct $z \in A(t)$.*

It is worth emphasizing that empirically, recovering this type of data is much more demanding than from a retrospective perspective. In particular, while the retrospective model involves the elicitation of preferences at one period of time, the evolving model entails the elicitation of identification volitions before each period.

Note that the notion of second-order volitions corresponds, in its most basic sense, to a theory of preference formation and change. It would be then possible to impose an axiomatic structure that would require that choices are the result of preferences motivated by second-order volitions. For example, whenever these volitions would move an agent towards a change in preferences, an evolving model could require such change in preferences to occur and the potential subsequent choice reversal to follow. This would be in line with the economic tradition of identifying the observable choice implications of different decision making models, as it is explored

in Chapter 3 of this dissertation. It is worth noticing, however, that such formulation would be somewhat contrary to the philosophical view of second-order volitions. Even though their feature of commitment, an agent may fail to follow the decision he has taken in accordance to a second-order volition due, for instance, to a strong and irresistible desire to choose otherwise (see e.g. Frankfurt 1988). A non-deterministic view of agency, as alluded in the conclusion of this chapter, could potentially explain and rationalize this kind of behavior. In any case, the development of such formal framework is a possible venture left for future work.

1.6 The Reflexive Agent

The possibility of reflexive preference change suggests the relevance of models that incorporate and explain changing (first-order) preferences. Decision making models based on multiple preferences are a possibility. They have shown to be a successful instrument to provide rationalizations for changing preferences and choice heuristics that may be behind some of the cyclical patterns of choice observed in real data (e.g. Aizerman and Malishevski 1981; Manzini and Mariotti 2007; see also the changing preferences model of Chapter 3). One way to represent a reflexive agent is then through multiple preferences that she alternates over time *and* identify or not with. This presupposes combining multiple preferences models with an explanation of reflexive preference change, such as the hierarchical evolving model just described.

Observationally, while the multiple (first-order) preferences may be recovered through choice data, data on second-order preferences may be collected, notably, from non-choice data such as individuals' verbal evaluations of their preferences/choices. In Section 1.8.1 I return to this topic and briefly discuss one survey-based and one choice-based method of eliciting second-order preferences.

In what follows, I discuss two representations of the economic agent that can be associated with a hierarchical (evolving) model in order to model reflexive and non-reflexive preference change. One is based on the conflict between an individ-

ual's multiple preferences (Section 1.6.1). The other is based on the evolution of an individual's single preference (Section 1.6.2).

1.6.1 The Conflicted Agent

> "People behave sometimes as if they had two selves, one who wants clean lungs and a long life and another who adores tobacco, or one who wants to improve himself by reading Adam Smith's theory of self-command (in *The Theory of Moral Sentiments*) and another who would rather watch an old movie on television." (Schelling 1984, 58)

The conflict between different preferences is a topic that has been widely discussed and modeled in economics. Models in this tradition represent the economic agent as if endowed with a collection of preferences, i.e., a plurality of distinguishable identities, roles, motivations, or points of view that are fixed over time. In some cases these are represented as a collection of orderings, and others as multiple *selves* (or "subagents") that interact with each other as if they were players in an interpersonal game. This representation, as noted by Gul and Pesendorfer (2008, 30), represents a "departure from the standard economics conception of the individual as the unit of agency".

This representation is in tune with views such that of Schelling (1984), who considers that people are best represented as a collection of "values centers" that share the same beliefs and reasoning capacities but differ in terms of volitions. According to his view, one value center (or self) will act as if a dictator at each period, winning "the intimate contest for self-command" at that period (see Schelling 1984, 57-81). From period to period, individuals "alternate" from one preference to another. An alternative and recent example is given by the *reason-based* theory developed by Dietrich and List (2013, 2016), in which an agent's preferences over alternatives depend on her *motivational state*, defined as a subset of all possible *motivationally salient* properties of those alternatives (the properties that the agent focus on). Then, in their

model, an agent is represented as a family of preference relations over all possible motivational states and alternates from one preference relation to another according to the motivational state in which she happens to be.

Kalai, Rubinstein and Spiegler (2002) explore a testable model consistent with this view, in which an agent chooses the best alternative according to one of multiple preferences in each choice situation. Choice behavior is rationalized by a collection of preference relations, such that for every choice situation A in the domain of feasible choice situations, the chosen alternative is maximal in A for some preference in the collection. Note that with such a decision making model, one can always rationalize any choice behavior whatsoever by resorting to a sufficiently large collection of different preferences. Still, Kalai et al. (2002) show that a plausible upper bound on the number of different preferences that can rationalize a choice behavior generated by a universe containing n alternatives is $n - 1$. More recently, Apesteguia and Ballester (2010) have built upon this framework and studied the complexity of finding the minimal number of preference relations - the lower bound - necessary to explain choice behavior.

Conflicted agent models could possibly explain the preference changes/reversals discussed earlier. If an individual identifies with her multiple preferences, and alternates between those she identifies when choosing, then the behavior of this person may be inconsistent with the standard revealed preference axioms without being interpreted as a mistake. If, instead, the individual does not identify with some of her multiple preferences, then some changes in the preferences exhibited by the choices of this person may not be the result of reflexive preference change. To distinguish reflexive from non-reflexive preference change, conflicted agent models need be coupled with an hierarchical model in order to specify if, at any given point in time, the multiple preferences and identification preferences of the agent are aligned.

In addition, conflicted agent models may be reasonable approximations of the mode of *reasoning* of individuals, or, at least, a convenient way of describing human psychology. One important distinction is between models that assume that the

conflict is the result of the switch among a collection of preferences (as e.g. Kalai et al. 2002 and Dietrich and List 2013, 2016), and the models that assume that the conflict is the result of the strategic interaction between multiple selves modeled as "subagents" (as e.g. Schelling 1984). As for the former, they seem useful to study the behavior of individuals that switch preferences according to the *role* they happen to be playing or according to the (social) identity or motivation that happens to be salient at a given point in time. Recent experiments suggest that this may be a meaningful exercise. For example, in the case of Asian-American subjects, some experiments suggest that it is sufficient to make one preference (identity) more salient than another (the Asian or the American identity) to trigger different behavioral responses in terms of patience (Benjamin, Choi and Strickland 2010) or cooperation (LeBoeuf, Shafir and Bayuk 2010).

As for the later, the analogy between interpersonal and intrapersonal conflict is useful since economists have developed a wide range of tools to study interpersonal conflict that can, in this way, be used to study cases of intrapersonal conflict. However, it is worth noticing that the analogy between intrapersonal and interpersonal conflict may be sometimes misleading. For instance, "[p]eople are able to punish or control each other to avoid conflict in a way that is not possible among 'multiple selves' " (Arlegi and Teschl 2015). According to Elster (1986, 2) "the possibility of mutual strategic interaction [...] is hardly plausible", and *deception* and *manipulation* may be the essential forms of interaction. And despite neurological evidence that the brain may sometimes work in this fashion (see e.g. Jamison and Wegener 2010), it seems that we still need to understand when such representation is an accurate description of *why* people make their decisions. Only with correct underlying assumptions about the individuals' motivations and mode of reasoning are we able to fully understand cause-and-effect relationships, and make predictions that will remain correct throughout different environments (see Schotter 2008, 72).

Finally, one can remark that the change in preferences in a conflicted agent model is made from a *fixed* collection of preferences. In a dynamic perspective, this means

that the possible preferences that the agent can hold do not change over time. This is unlikely to hold in cases in which a person's values/second-order preferences change over time. Next, I discuss a representation of the economic agent that is in tune with the endogenous change of preferences as the ability to evaluate and change one's single preference, one's identity, or *who* one is or want to become.

1.6.2 The Evolving Agent

> "Thus we say of an oak that it is the same thing from the seed to the tree in the prime of life. The same is true for an animal from birth to death, and for a man, as a specimen of the species, from foetus to old man. The demonstration of this continuity functions as a criterion supplementary to that of similarity in the service of numerical identity. The contrary of identity taken in this third sense is *discontinuity*. But what has to be taken into account in this third sense is change through time." (Ricoeur, 1991, 190)

An alternative representation of the economic agent is to assume that the agent is endowed with one personal identity that evolves over time. This conceptualizes the individual as an *evolving agent* that makes her decisions according to an endogenous sequence of (multiple) preferences. Contrary to the rational agent model that assumes a stable personal identity over time, and differently than the conflicted agent model that assumes independent and conflicting preferences or selves at each period, the evolving agent model is a representation of the economic agent with a single preference that evolves over time.

While in economics, to the best of my knowledge, the endogenous evolution of a single preference is not a representation that is often used, the evolving agent model is consistent with a notion of the self (or personal identity) that finds support in philosophy, psychology, and neuroscience: the narrative identity or the narrative self. The "narrative self" is defined as a "more or less coherent self (or self-image)

that is constituted with a past and a future in the various stories that we and others tell about ourselves" (see Gallagher 2000, 15). According to Paul Ricoeur (who conceptualizes this notion e.g. in 1984 and 2002), a person acts according to a personal identity that extends *and* evolves over time.[26] Ricoeur draws attention on the fundamental distinction between two uses of the concept of identity: identity as sameness (*idem*) and identity as self (*ipse*). While *idem* refers to a notion of identity of something that is always the same, immutable, permanent and unified, *ipse* refers to a notion of identity of one's selfhood through time and change. While the (inner) rational agent rests on the *idem* notion of identity, the evolving agent can be connected to both notions: I am and I am not *who* I was five years ago. Dennett (1991) proposes a version of this concept in neuroscience in which the self is defined as a "center of narrative gravity", where the various stories told about the person meet (see also Gallagher 2000). Albeit the narrative self is, in this case, seen as a fictional representation, it is an important principle of organization that through various narratives makes one's experience relatively coherent over extended periods of time. Changes in preferences, in this perspective, are mediated by the way one recounts to oneself the experiences that one has lived and is to live.

While the representation of an evolving agent points towards an endogenous sequence of multiple preferences, it is observationally consistent in terms of choice with an exogenous sequence of independent preferences or selves. Such representation has been sometimes used in economics. For instance, Gul and Pesendorfer (2001, 2004, 2005) model of intertemporal choice and the changing preferences model of Chapter 3 of this dissertation are consistent with the exogenous change of preferences. Similarly, Strotz (1955) interpretation of intertemporal choice illustrates an extreme version of this type of model: Instead of two preferences or selves that are in conflict at each point in time, the conflict is between today's and tomorrow's pref-

[26] See e.g. Davis (2009) for another theory of personal identity as narrative identity. See Kirman and Teschl (2006) for a social identity perspective on the evolution of one's identity that depends on what a person currently is and does, *who* she wants to be or become, and to which social group she chooses to belong (see also Horst, Kirman and Teschl 2006).

erences. According to Strotz (1955, 179), "[t]he individual over time is an infinity of individuals".

An important feature of the evolving agent model is that it is consistent with a *person*'s reflexive capacity to evaluate and change her identity or *who* she is, wants to be or become. As argued in the previous sections, the evolution of a person's preferences may be the result of the resolution of a conflict between first- and second-order preferences. I may, upon my experience, form a second-order preference (an identification preference) or volition against a preference that I currently hold, and in the future change my preference accordingly. But at each point in time, an individual may or may not identify with one's present preferences, as well as with the previous or next preference change. I may, for instance, not identify with a preference change or choice reversal that I feel I cannot avoid, as in the cases of relapse into addiction. The evolving agent model would accommodate, in this sense, the reflexive and non-reflexive evolution of a person's single preference.

In order to distinguish between reflexive and non-reflexive preference change, the evolution of first-order preferences must be accompanied by the evolution of the agent's second-order preferences as, for example, in the hierarchical evolving model above. In the case of models of exogenous preference change, the collection of non-choice verbal data on second-order attitudes such as identification preferences could inform if such apparently exogenous changes of preferences are the result, at least in retrospect, of reflexive or non-reflexive preference change. More structure would need to be imposed over the (hierarchical) evolving model in order to explain or predict some patterns of behavior such as the repetition of certain bias over time.

The evolving agent model seems philosophically and psychologically appealing. It may also help in centering economic analysis around meaningful questions, in particular on frequent and reasonable behavior due to preference change. The evolving agent model is not only compatible with learning and preference formation, but, also, with changing tastes or values that depend on the experience of the economic agent. Despite the appeal of the evolving agent model, one finds several economists

who argue that for most purposes stable preferences are a necessary requirement for economic analysis. For instance, Samuel Bowles (1998, 79), in a survey about endogenous preferences, writes the following:

> "For preferences to have explanatory power they must be sufficiently persistent to explain behaviors over time and across situations. If preferences are endogenous with respect to economic institutions it will be important to distinguish between the effects of the incentives and constraints of an institutional setup (along with given preferences) on behaviors, and the effect of the institution on preferences per se. The key distinction is that where preferences (and not just behaviors) are endogenous they will have explanatory power in situations distinct from the institutional environments which account for their adoption. Thus, however acquired, preferences must be internalized, taking on the status of general motives or constraints on behavior. Values which become durable attributes of individuals - for example, the sense of one's own efficacy introduced below - may explain behaviors in novel situations, and hence are included in this broad concept of preferences."

The "broad concept of preferences" defended by Bowles (1998) conceptualizes preferences as attributes that are (endogenously) "acquired", but that "become permanent reasons for behavior" (Bowles 1998, 80). According to the author, only such preferences have explanatory power. Similarly, Hoff and Stiglitz (2016, 2) stress that "[p]ast social experiences and past social structures can result in sustained ways of conceptualizing a situation and, hence, sustained beliefs and social outcomes". The authors propose to focus on an "enculturated actor", whom preferences, perception, and cognition are subject to "deep" social influences rooted on the social and cultural backgrounds he is exposed to. Indeed, the dependence of preferences and behavior on "deep" social influences seems quite sensible, and a lot of empirical findings, part of them reviewed in Bowles (1998) and Hoff and Stiglitz (2016), point that this is

indeed the case. I side with these authors that the endogenous determinants of preferences should be central in economics (as Chapter 5 of this dissertation testifies).

However, this is not incompatible with reflexive and non-reflexive preference change. Besides understanding the cultural and social determinants of preferences, it seems important to understand what might *change* these preferences. Market rules and economic institutions evolve over time. It is only by properly understanding why preferences change that we will be able to appraise the effect(s) of economic policies and changes in economic institutions on preferences. Carpenter (2005) provides an example. The author conducted a within-subject experiment to test if individuals' social preferences change according to different aspects of the market. His findings suggest that subjects are less pro-social in more anonymous settings due to a change in preferences. Since social preferences are likely to reflect, in some part, one's social values, it seems that these changes can be, at least in principle, the result of reflexive preference change.

In addition, it seems possible to make meaningful economic analysis even when preferences are not durable attributes of individuals. To make this point, it is useful to consider the model of changing preferences of Chapter 3. In this model, an agent *may* change preferences from one period to another, and we analyze the case in which the agent preferences may change unpredictably *at most* once. We show that an analogous condition to GARP is necessary and sufficient for the behavior to be explained by the maximization of a single preference relation between any two (not necessarily consecutive) periods, and that to rationalize one change in preferences one can apply GARP in two partitions of the time horizon. Thus, if one observes a violation of GARP between a period 1 and a given period t, the observable condition that rationalizes one change in preferences says that it is not possible to observe a second violation of GARP between t and the last period of the time horizon. Note that one could use a similar reasoning to rationalize more than one change in preferences. Then, one can assume that the behavior that satisfies GARP between any two (not necessarily consecutive) periods is the result of the maximization of a single

preference. In this context, stability is not an *a-priori* assumption, but an observable and verifiable condition for any sequence of periods. In the case it is verified for any sequence of periods, it is then possible to use the consistency properties associated with stable preferences for this sequence of (time) periods. This means that the evolving agent model allows, at least in principle, to use with considerable generality and tractability many of the tools economists have developed so far.

This example illustrates that an evolving agent model can be empirically refutable. An evolving agent model would not be empirically refutable in terms of choice if no *a-priori* restriction on the number of changes of preferences is imposed, since any choice behavior whatsoever is rationalizable if we assume that the economic agent is choosing according to a single preference (possibly distinct) at every period. An evolving agent is, in this sense, trivially consistent with time- and context-dependent preferences. But as Chapter 3 of this dissertation illustrates, it is possible to impose meaningful restrictions upon choice behavior when one predicts that a limited number of changes in preferences are to occur. For instance, the changes in preferences following a change in the market rules or conditions may be anticipated and modeled as a single event of *potential* change in preferences. This means that an observer could assume, *a priori* and if desirable, that the economic agents have stable and context-independent preferences prior and after the change in the market. Brennan (1993) makes a similar point, arguing that it is worth studying preference change whenever such change is predictable:

> "Nothing in an SU [single utility] theory rules out preference changes. I will grant Professor Lutz that most economists assume preferences are stable. This assumption might be defensible because it deters us from too quickly invoking preference change to salvage a failed prediction and encourages us to base our explanations of behavior on observable phenomena such as prices or incomes. Of course, this defense can turn into demagogy. Preference change may be real and perhaps even predictable.

> I have elsewhere acknowledged that where preference change is likely, as in broadcasting, efficiency models may make no sense (Brennan, 1983). In such contexts, I agree with socioeconomists that economic explanations and policy evaluations should be supplemented by models of preference change." (Brennan 1993, 162)

Finally, the evolving agent model opens the possibility to search for the lower or upper bound on the number of preferences necessary to explain choice behavior. As discussed above, Kalai et al. (2002) and Apesteguia and Ballester (2010) show that this is a meaningful exercise for a model that is, in fact, observationally consistent in terms of choice with an evolving agent. Then, in cases of unpredictable number of changes in preferences, one could search for the minimal number of preference relations necessary to explain choice behavior within a given sequence of (time) periods.

1.7 Preferences over Preferences in Welfare Economics

> "I merely wish to emphasize here that we must look at the entire system of values, including values about values, in seeking for a truly general theory of social welfare". (Arrow 1951, 18)

One important aim of welfare economics is to provide rankings of individual and social welfare.[27] In neoclassical economics, preference satisfaction is one of the main, if not the dominant view of individual and social welfare. Even when confronted with the evidence that observed behavior differs from the maximization of a stable and context-independent preference, economists often take the satisfaction of a *given* preference relation as the benchmark for welfare analysis (e.g. Koszegi and Rabin 2007; Rubinstein and Salant 2012; Apesteguia and Ballester 2015).[28]

[27] Throughout this section I will discuss how different welfare criteria may or may not apprehend individuals' welfare, and how they *can* be useful for an observer (e.g. planner) to make inferences about individual and social welfare.

[28] See Infante et al. (2016) for a review and criticism of this approach.

Apesteguia and Ballester (2015) approach to welfare illustrate this well. In their framework, an index indicates how "far" choice behavior is from the maximization of a single preference relation. Their index ranks different preference relations according to the number of alternatives in each choice problem (among the available ones) that need to be "swapped" with the chosen alternative in order to rationalize individual choices. Then, relying on standard revealed preference data, they interpret the preference ordering that "minimizes" that index (i.e., the unobservable preference relation that is "closest" to the revealed choices) as a reflection of the higher attainable individual welfare. In this sense, any decision that is inconsistent with the traditional assumptions about "rational choice" is seen to entail a welfare loss. Another example is given by Rubinstein and Salant (2012), who assume that an agent reveals distinct preference relations in different contexts that vary according to properties of the choice environment that are deemed *normatively irrelevant* (e.g. frames). They assume that the multiple preferences are the outcome of some cognitive process that distorts the agent's unobservable and underlying preference that reflects her welfare. The authors then define testable assumptions on the decision process that relate unobservable preferences to choice behavior in order to elicit the agent's underlying preference. As in the standard neoclassical approach, the satisfaction of some *context-independent* preference, in these papers *purified*[29] of mistakes, is used as a normative criterion.

Although they do not assume the existence of underlying preferences, Bernheim and Rangel (2007, 2009) build a choice-theoretic welfare criterion that relates with this view.[30] They wish to respect individuals' choices in the presence of context-dependent behavior. In order to handle such context-dependent behavior, the authors propose a *generalized choice situation* $GCS = (A, d)$ where A corresponds to a standard choice situation and d to an *ancillary condition* such as the manner in which information is presented or other frames. Like Rubinstein and Salant (2012),

[29] I borrow this term from Hausman (2012) and Infante et al. (2016).

[30] See also Fleurbaey and Schokkaert (2013) who introduce interpersonal comparisons and distributional considerations within a framework that extends Bernheim and Rangel (2007, 2009).

they deem these ancillary conditions as normatively irrelevant, i.e., "a feature of the choice environment that may affect behavior, but [that] is not taken as relevant to a social planner's evaluation" (2009, 55). They then define a welfare criterion based on what they call an "unambiguous choice relation", for which x is said to be unambiguously chosen (and welfare superior) over y if and only if y is never chosen when x is available. In this sense, only context-independent choices (that remain stable for all ancillary conditions) are considered to reveal the individuals' welfare relation. From a revealed preference perspective, this corresponds to the satisfaction of some context-independent preference relation.

The recognition that preferences change over time highlights one of the difficulties with taking the satisfaction of a given (revealed) preference relation, even when "purified" of supposed mistakes, as a measure of well-being and social welfare. Which preferences, from the several that are one own over time, should be given priority? Should or should not an observer dismiss a preference that was revealed by an agent in the past but that she no longer holds?[31] And, as Hausman (2012, 81) rightly asks, "should the consequences for welfare of a policy that changes people's preferences be measured by people's preferences before the policy is put into place or by their preferences afterward?"

Take the example of Bernheim and Rangel's (2007, 2009) welfare criterion with the ancillary condition of a time horizon $d = \{1, ..., T\}$. In their framework, if x is chosen over y at period $t-1$ and y is chosen over x at period t then x and y are said to be non-comparable in terms of welfare. In this sense, "past" $(t-1)$ and "present" (t) choices (and the potential preferences behind them) are taken *equally* in consideration. However, in many cases our intuition seems to suggest that time is normatively relevant. For instance, it seems odd to give normative authority to my childhood's preference to be a writer of poems (see Parfit 1984, ch. 8). Particularly, since today I do not identify with and/or care about being a writer of poems. More generally, it seems quite plausible to say that "[a]nyone can rationally ignore the

[31] See e.g. Parfit (1984) and Bykvist (2003) for critical analysis of these questions.

desires that he lost because he changed his mind [desires that are no longer judged of worth or important]" (Parfit 1984, 153).

More generally, the arguments in the preceding sections suggest that it is important to distinguish between reflexive and non-reflexive preference change. As argued before, this is possible by taking individuals' second-order preferences (when seen as identification and valuation preferences) into consideration. Whenever preferences (or choices) are not consistent with the standard neoclassical representation, identification preferences can bring information that allows one to infer if such "inconsistencies" result from reflexive or non-reflexive preference change. In the case of the childhood's preference, it seems uncontroversial that if the individual no longer identifies with such preference that this has been a reflexive preference change. In this sense, second-order preferences inform when to ignore past desires.

Valuation preferences can bring information on how a person reflexively ranks her choices in terms of value. In particular, valuation preferences can reflect *an order* of what people care about. That is, they can bring ordinal information on what an individual *values*, what is *important* to her, or what she *cares* about. This information is relevant, among other things, because "[b]esides wanting to fulfill his desire, [...] the person who cares about what he desires wants something else as well: he wants the desire to be sustained" (Frankfurt 2009, 16).

Second-order preferences may also enrich normative frameworks with forward-looking information on what people want to achieve and *who* they want to be or become. A preference (choice) that is high ranked in a person's valuation preference and that the individual does not identify with is an indication that the reversal of this preference (choice) might be an important goal for the person (who the person wants to become). And a preference (choice) that is high ranked in a person's valuation preference and that the individual identifies with is an indication that this preference (choice) might reflect who the person wants to be. As argued by Kirman and Teschl (2006, 319), "[t]he extent that a person manages to become and to be *who* she wants to be can be said to be a particular measure of her well-being and quality of

life." Second-order preferences, upon some index transformation, could potentially provide such kind of "measure".

According to these arguments, second-order preferences may enrich the neoclassical approach to welfare since revealed preferences alone will often fail to reveal what constitutes welfare and what is important to people both in the present and in the future. In addition, second-order preferences (when seen as identification and valuation preferences) may have several side advantages such as "purifying" choices/preferences from the individual's *own* perspective. For instance, some identification preferences that contradict first-order preferences may refer to first-order preferences based on expensive tastes, *antisocial desires*[32], manipulation or coercion. I can have a taste for champagne, but would prefer not to prefer to have this taste. This seems to bring more confidence to an observer that, presented with the task of deciding between subsidizing two substitute goods on a limited budget, one non-expensive (say sparkling wine) and another expensive (say champagne), to subsidize the former. In this sense, the information on identification preferences would allow an observer to make an informed decision that would respect a person's evaluation about herself even if this evaluation has not caused action. Another side advantage of using second-order preferences is connected with the process of obtaining the necessary information. By asking people to state the evaluations of their own preferences (choices), the observer is letting people to think about their preferences (choices) and might led them to choose in future occasions (by their will and after a potentially slower deliberation) the option(s) that the observer *a priori* considers most favorable to them.

For example, a potential application is to use second-order preferences as a refinement to welfare rankings that combine (i) utility-based notions of well-being with (ii) survey-based data (e.g. Benjamin, Heffetz, Kimball and Szembrot 2014). Adding counterfactual questions regarding "what a person would prefer herself to choose" (or prefer) can inform not only on issues of self-control (as in Benjamin,

[32]Defined as intrinsic preferences against the well-being (or freedom) of another.

Heffetz, Kimball and Rees-Jones 2012), but also on the values and goals of a person for herself (and possibly others). Questions concerning valuation preferences may bring novel information on the priorities of people over what they care and want to be or become. This would partly countervail one of the two main criticisms posed by Sen (1987, 14) against the traditional utility-based notions of welfare (either as happiness or desire-fulfillment), that "have the twin characteristics of (1) being fully grounded on the mental attitude of the person, and (2) avoiding any direct reference to the person's own valuational exercise - the mental activity of valuing one kind of life rather than another."

1.7.1 Individual Sovereignty, Opportunities, and Context-dependency

To substantiate the usefulness of second-order preferences, I would like to discuss two criticisms of their use put forward by Robert Sugden (2004). Though they are addressed to the traditional meta-ranking view of second-order preferences, it is equally pertinent to discuss them for the present approach. The two criticisms are framed as critics to second-order preferences as an alternative approach to the author's "opportunity criterion" that is intended to respect individuals' choices without referring to the preferences that lie behind them. The first criticism is that meta-rankings based on second-order preferences are opposed to the value of opportunity and individual sovereignty. As the argument goes, "[t]he metaranking approach locates normative authority, not in the day-to-day decisions that individuals make as economic actors, but in each person's supposed higher moral self" (Sugden 2004, 1017). A robust concept of individual (consumer) sovereignty, according to this view, "should not need to invoke such a moralized account of preference". And since a second-order preference may be contrary to a first-order preference (i.e., it may value it negatively), the second-order preference may be read as a prescription for imposing restrictions on the individual's opportunities to fulfill such first-order preference.

First, I would like to argue that part of this criticism is at odds with Frankfurt's (1971, 13) view that second-order preferences/volitions are not necessarily moral:

> "In speaking of the evaluation of his own desires and motives as being characteristic of a person, I do not mean to suggest that a person's second-order volitions necessarily manifest a moral stance on his part toward his first-order desires. It may not be from the point of view of morality that the person evaluates his first-order desires. Moreover, a person may be capricious and irresponsible in forming his second-order volitions and give no serious consideration to what is at stake. Second-order volitions express evaluations only in the sense that they are preferences. There is no essential restriction on the kind of basis, if any, upon which they are formed."

According to this view, second-order preferences do not necessarily represent a "higher moral self", but instead the relative importance that a person gives to different first-order preferences that she holds. More specifically, and as argued above, second-order preferences can represent - with the exception e.g. of times when a person gives "no serious consideration to what is at stake" - what a person values and cares about. For instance, I might would like to quit smoking (even if my preference is for smoking), not because I find it the moral thing to do, but because I care about my health *or* simply because it is economically-wise. In this sense, second-order preferences are relevant not because they are moral but because they *often* reflect what a person values or cares about.

Second, I wish to argue that it is possible to take second-order preferences into consideration and still respect the value of opportunity and individual sovereignty. Take the example of Elizabeth, a smoker who has stated that would like to quit smoking (and that cares about quitting smoking). Given that preferences in general determine choices, she thus *reveals* a first-order preference for smoking, and *states* a second-order preference not to smoke. Note that this is the canonical example of a

second-order preference that contradicts a first-order preference. Following Sugden (2004), an observer who wishes to respect the value of opportunity and her individual sovereignty should not prohibit smoking. This would reduce her opportunities and be against her first-order preferences. I agree.

Still, why should not Elizabeth's evaluations about her preferences be important when considering her sovereignty? Why are her revealed preferences more important than her evaluations about her preferences in this respect? Although the arguments above suggest that sometimes one should give priority to second-order preferences over first-order preferences, I wish to argue that it is possible to respect Elizabeth's opportunities and the sovereignty of her first-order preferences even when taking second-order preferences into consideration. Indeed, it is even possible to enhance Elizabeth's opportunities and her sovereignty when opportunity and sovereignty are seen from a broader "positive" perspective. To do so, it is convenient to recall the distinction between *negative* and *positive* freedoms. In its most famous formulation, due to Berlin (1969), the former is interpreted as the freedom from constraints that are imposed by others (as opposed to constraints such as economical or biological impediments).[33] For Berlin (1969, 8), negative freedom consists in "not being prevented from choosing as I do by other men". Positive freedom, still according to Berlin (1969), consists in the ability to lead a life in an autonomous and reasoned/conscious fashion, in a way that actions are *self-directed* and not influenced by external nature and other man. In the words of Berlin (1969, 8), "[t]he freedom which consists in being one's own master". In economics, Sen (1988) favors a notion of positive freedom that consists in what a person is able to choose to do or achieve. This corresponds to focus on the actual or "real" opportunity to choose, rather than on the absence of constraints to achieve certain goals.

[33]The nature of the relevant constraints for the negative notion of freedom has been the subject of many debates on political theory. In classical debates this notion was associated with political freedom (the limits on free action that should be imposed by law) and the frontier *between the area of private life and that of public authority* (see Berlin 1969).

Now, if we interpret respecting Elizabeth's opportunities and sovereignty in *positive* terms (i.e., including both negative and positive freedoms), it is possible to take second-order preferences into account and respect or even enhance, in this sense, her opportunities and the sovereignty of her preferences. In particular, if an observer wishes to respect the sovereignty of Elizabeth's first- and second-order preferences they should not only *not* prohibit smoking *but also* enhance some policy that would help her to surpass what *she* considers to be a negative preference/behavior (or her "weakness of will"/addiction, since her behavior is not consistent with her will). For instance, providing Elizabeth with free consultation(s) with a specialized doctor could be such a policy. By doing so, the observer would still respect the value of opportunity (in terms of negative freedom), while in addition enhancing her positive freedom to lead the life that according to *her* second-order preferences she has a reason to value. Without such policy, Elizabeth has the negative freedom to go to a consultation (it suffices that it is available on the market), but she may not have the positive freedom (or "real" opportunity/"capability") to do so, for, say, budget constraints. Since the two policies (non-prohibition and free consultation) are not mutually exclusive, the observer would *respect* Elizabeth's sovereignty in terms of her first- and second-order preferences, which seems to favor a broader/positive view over individual sovereignty.

The second criticism of second-order preferences (but also of preference satisfaction in general) presented by Sugden (2004) is based on the evidence that preferences are often susceptible of changing according to *trivial changes in viewpoint or context*. As the argument goes, as soon as we acknowledge that preferences are unstable or context-dependent, using preference satisfaction (either of first- or second-order) for normative analysis is not possible:

> "Economists often want to make normative comparisons between very different social states - for example, between a future in which international trade is subject to tariffs and one in which it is not. The standard meth-

> ods of welfare economics hold individuals' preferences constant across the relevant social states, treat those constant preferences as measures of well-being, and ask how far they are satisfied in each state. Such analysis is not possible if individuals' preferences shift according to trivial changes in viewpoint or context." (Sugden 2004, 1016)

Sugden (2004) is right in arguing that if preferences are totally contingent upon *trivial changes in viewpoint or context*, then normative analysis is better done without taking preferences into consideration. Still, there seem to remain cases in which preferences change for predictable and reasonable reasons. As argued in Section 1.6.2, preferences may change after some predictable or known event; they can also change because of the resolution of a conflict between first- and second-order preferences (e.g. Elizabeth adopts a first-order preference not to smoke, and consequently stops smoking).

In addition, as the literature on the measurement of freedom and the ranking of opportunity sets illustrates (e.g. Pattanaik and Xu 1990; Foster 1993; Gravel 1994; Nehring and Puppe 1999)[34], excluding information on individual preferences may leave us with coarse criteria. In the setting of ranking opportunity sets according to the freedom they offer, such an approach often leads to rankings based on the number of alternatives of each set (e.g. Pattanaik and Xu 1990). However, it is questionable that a set containing two "good" alternatives provides the same amount of freedom than a set containing two "bad" alternatives.[35]

Despite these arguments in favor of using preference information, I side with Sugden (2004) in that objective measures of well-being (such as opportunities) are necessary and should be central in welfare analysis (see also Section 1.8.4 below). Preferences (either of first- or second-order) are based on what people want, their

[34] An opportunity set is any set of alternatives (assumed to be mutually exclusive) that are available for choice for an individual. The main question in this setting is what it means - according to a definition of freedom or opportunity - for one opportunity set to offer more freedom/opportunity than another. See Gravel (2008) for a comprehensive review of the use of the notion of freedom in economics.

[35] This discussion is often centered around one of the axioms of Pattanaik and Xu's (1990) ranking of opportunity sets that states that all opportunity sets that contain one alternative (singletons) offer the same amount of freedom of choice. See Jones and Sugden (1982) and Sen (1991) for competing views.

desires and goals. It follows that alternatives that are not the focus of the desires, wants, or goals of individuals have no normative authority. Then, as pointed by Sen (1980, 210), "[o]pportunities have no value in a desire-supported system, only *desires* for opportunities have, and objective contraction of opportunities can be washed out by subjective change of desires." Further, I believe that Sugden's (2004) arguments bring more strength to this view. Given the potential contingency of preferences (either of first- or second-order) on trivial changes in viewpoint or context, it will sometimes be difficult to have credible rankings based on preference information. Nonetheless, what I intend to suggest in this section is that we may not want to forget about the information on preferences all together, but try instead to have richer data sets that include information on first- and second-order preferences. I turn now to a potential application that would use a data set enriched in this sense.

1.7.2 An Application to Bernheim and Rangel (2007, 2009)

Information on second-order preferences can be used as what Bernheim and Rangel (2007, 2009) call a *refinement* of their welfare ranking. As pointed by Rubinstein and Salant (2012) and others, Bernheim and Rangel (2007, 2009) welfare ranking is typically a coarse binary relation that becomes more so as the number of choice observations increases. For that reason, Bernheim and Rangel (2007, 2009) propose to use "nonchoice evidence [such as evidence on inattention] to officiate between conflicting choice data by deleting suspect GCSs" (2007, 469). In this section, I use the retrospective model and identification preferences to select "suspect" $GCSs = (A, d)$.[36] The feature of the choice environment d is then the periods at which the choices have been made. In the framework of the retrospective model, Bernheim and Rangel's (2009) preferred welfare criterion can be written as follows:

[36]For the sake of presentation I abstract from valuation preferences, but they could bring meaningful information, for example, in terms of priorities over policy measures associated with this kind of welfare ranking.

Definition 8. *Alternative x is said to be **welfare choice-superior** to alternative y if and only if $y \neq C(t, A(t))$ for all $A(t) \in X$ such that $x, y \in A(t)$.*

Identification preferences provide a non-arbitrary justification to exclude GCSs based on the individuals' *own* evaluations/judgments of their choices. It is also reasonable to take them into account if one wishes to respect individual sovereignty in the broad (positive) sense defended here. For example, if when provoked I harmed another man when I would have preferred to stay calm, it is suspect to consider that the choice of harming the other man (as opposed to staying calm) was the best in terms of my well-being. Similarly, if between two cellphone alternatives I buy an expensive one (as opposed to an economical one) but asked about this choice I state that it was a mistake (because, say, I value and identify myself with being frugal but was unable to be so because of the beautiful commercial add in favor of the expensive cellphone), it seems at least prudent not to consider this choice to reflect what is best for my well-being.

In order to respect individual sovereignty one would then delete (or not) GCSs based on the agent's identification preferences. As an illustration, take the example of binary choices. One can then distinguish two cases. The first is when an individual does not identify with a given choice (i.e., the choice is a self-authenticated mistake). In this case, even if x is *consistently chosen* over y there is an argument to not count x as being welfare superior to y. Notice that this is a prudent refinement that would turn a welfare ranking based on GCSs *coarser*[37]. The second is when an individual is inconsistent in her choices but identifies with one choice (say of x over y) and not with the other (of y over x). In this case, even if y has been chosen over x, there is an argument to count x as being welfare superior to y. Notice that in this case this refinement would turn the welfare ranking *finer*. An example of the second sort of cases is when an agent identifies with preference change. For instance, an agent loves meat but becomes a vegetarian for ethical reasons. At period T, she does not

[37] A ranking becomes coarser (finer) when it becomes less (more) discerning. Formally $\succsim$ is said to be coarser than $\succsim'$ if $x \succsim' y$ implies $x \succsim y$, and if $\succsim$ is coarser than $\succsim'$ then $\succsim'$ is said to be finer than $\succsim$ (see e.g. Bernheim and Rangel 2007, 469).

identify with her past choices of meat. Then, there is an argument to only take the later choice for vegetables into account even if meat has been chosen in the past.

Since we are looking at choices over subsets of any size, it is possible to have many different welfare criteria that take identification preferences into account. I give two examples based on the definitions introduced in Section 1.5: One that respects reflexive choices, and another that discards mistakes. The former can be stated as follows:

Definition 9. *Alternative x is said to be **welfare reflexive-superior** to alternative y if and only if (i) $x = C(t, A(t))$ for some $A(t) \in X$ such that $x, y \in A(t)$, (ii) $x \succsim_t^I w$ for some distinct $w \in A(t)$ and for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$, and (iii) there exists some distinct $z \in A(t)$ such that $z \succ_t^I y$ and/or there exist no $v \in A(t)$ such that $y \succsim_t^I v$ for all $A(t) \in X$ such that $x, y \in A(t)$ and $y = C(t, A(t))$.*

In words, x is said to be (welfare) reflexive-superior to alternative y if x is chosen at least once when y is present, if whenever x is chosen and y is present the choice is reflexive, and if whenever y is chosen and x is present the choice is *not* reflexive. Remark that this definition does not entail that for x to be welfare superior to y that x is *reflexive-consistently* chosen over y. It may be the case that y has been chosen for some $A(t) \in X$ such that $x, y \in A(t)$. The condition is that in that case the choice is not reflexive (but not necessarily a self-authenticated mistake). Remark also that in case of indifference in terms of identification over one's choice the criterion does not discard this choice. Take the example of binary choices. If y is chosen once over x and is indifferent in terms of identification to x in that period and x is chosen once over y and the agent identifies with that choice, then the ranking does not compare both alternatives in terms of welfare. Finally, note that this criterion, contrary to the one by Bernheim and Rangel's (2007, 2009), requires that $x = C(A(t))$ for some $A(t) \in X$ such that $x, y \in A(t)$ in order for x to be considered welfare superior to y. The underlying reason for this difference is that Bernheim and Rangel's (2007, 2009) preferred welfare criterion depends on defining the choice domain to include every

non-empty finite subset of X. In fact, if we do not observe an agent's choices from all these subsets (or at least from all pairs), as it will be most often the case, their preferred welfare choice-relation can be said to be less appealing. It could be the case that x would be said to be *welfare choice-indifferent* to y (as opposed to incomparable) even though x and y have never been compared in terms of choice.

I now state a criterion that instead of respecting reflexive choices discards mistakes:

Definition 10. *Alternative x is said to be* ***welfare self-superior*** *to alternative y if and only if (i) $x = C(t, A(t))$ for some $A(t) \in X$ such that $x, y \in A(t)$, (ii) there exist no distinct $w \in A(t)$ such that $w \succ_t^I x$ for all $A(t) \in X$ such that $x, y \in A(t)$ and $x = C(t, A(t))$, and (iii) there exist some distinct $z \in A(t)$ such that $z \succ_t^I y$ for all $A(t) \in X$ such that $x, y \in A(t)$ and $y = C(t, A(t))$.*

In words, x is said to be (welfare) self-superior to alternative y if x is chosen at least once when y is present, if whenever x is chosen and y is present the choice is not a self-authenticated mistake, and if whenever y is chosen and x is present the choice is a mistake. In the case of binary choices over x and y, this means that whenever y is chosen the agent identifies more with the choice of x. Remark, however, that the same is not necessarily the case for choices from subsets with more than two alternatives.

Under this general domain, these two welfare rankings (formulated in Definitions 9-10) are not necessarily complete, and more importantly, not necessarily acyclic. This may be problematic since without acyclicity it *may* not be possible to identify maximal alternatives for finite sets and/or unambiguous welfare improvements. Bernheim and Rangel's (2007, 2009) preferred welfare criterion (analogous to Definition 8) is also cyclic under this general domain. Their preferred welfare criterion is only acyclic when considering a choice domain that includes all conceivable choice problems of X. An open question remains concerning which restrictions upon the

binary relations and/or the choice domain would need to be imposed in order for the reflexive and self welfare rankings to be acyclic.

Finally, remark that identification preferences may be inconsistent over the same choice across time (e.g. $y \succ_t^I x$ and $x \succ_T^I y$). At T, I may say that when I was a child I identified with my desire to become a poet instead of a researcher (say $y \succ_t^I x$), while today I do not identify with such preference (say $x \succ_T^I y$). This unveils one potential issue with posing question concerning identification or of the type "what would you want yourself to have chosen?" when the time dimension is relevant. In this example, it seems that I have changed my values/second-order preferences, but it is possible that I would not want myself to have chosen differently in the past. In this kind of cases, the reflexive-choice criterion would deem the two alternatives as non-comparable in terms of welfare, when it seems that at least from the present perspective x is welfare superior to y.

One way to deal with these cases would be to, whenever identification preferences are not consistent over time, give priority to identification preferences over later (as opposed to former) choices. This could have the additional advantage of turning the welfare ranking finer. For instance, suppose that a person states that she identifies with the choice of x over y at period t, but she states that she does not identify with the choice of x over y from period $t+1$ onward. This seems to represent the childhood/adulthood preferences mentioned earlier. In this case, an observer interested in more discretion could take the later identification preferences as reflecting the (present) values of the person.

1.7.3 An Application to Intertemporal Preference Reversals

As a final illustration of the normative implications of reflexive preferences, consider again the case of preference reversals in intertemporal choice between a smaller short-term reward and a larger long-term reward. Consider the following example, taken from Gul and Pesendorfer (2008, 30-2). There are three consumption periods

($t = 1, 2, 3$), and three possible consumption paths (c_1, c_2, c_3): $(0, 0, 9)$, $(1, 0, 0)$, and $(0, 3, 0)$. In period 1, the agent chooses [prefers] $(0, 0, 9)$ over $(1, 0, 0)$ and $(1, 0, 0)$ over $(0, 3, 0)$. And in period 2, the agent chooses $(0, 3, 0)$ over $(0, 0, 9)$. Now suppose the agent faces the following decision problem: she can either choose $(1, 0, 0)$ in period 1 or leave the choice between $(0, 0, 9)$ and $(0, 3, 0)$ for period 2. *According to her preferences at period 1*, she chooses $(1, 0, 0)$. In fact, if she does not "commit" to this choice at period 1 and leaves the choice to period 2 between $(0, 0, 9)$ and $(0, 3, 0)$, she will end up choosing $(0, 3, 0)$ at period 2 which, although it is her most preferred option at period 2, it is the less preferred option from the point of view of period 1.

Gul and Pesendorfer (2008, 30), based on their previous work in Gul and Pesendorfer (2001, 2004, 2005), endorse a "standard, single-self model that accounts for this behavior". Denote by $\mathcal{C}$ the set of second-period choice problems, where $C \in \mathcal{C}$ consists of a consumption path with identical first-period consumption. Choosing $(1, 0, 0)$ in period 1 corresponds to $\{(1, 0, 0)\}$, while leaving the choice for period 2 consists of $C = \{(0, 3, 0), (0, 0, 9)\}$. Then, the authors describe period 1 preferences as follows:

$$\{(0, 0, 9)\} \succ_1 \{(1, 0, 0)\} \succ_1 C = \{(0, 3, 0), (0, 0, 9)\} \sim_1 \{(0, 3, 0)\}$$

where $\succ_1$ denotes a strict preference and $\sim_1$ indifference in period 1. This is consistent with the above preferences/behavior, since $\{(1, 0, 0)\}$ is ranked above the second period choice of $C = \{(0, 3, 0), (0, 0, 9)\}$. According to the authors, "[p]eriod 1 behavior reveals that the individual's welfare is higher in all periods when she is committed to $(0, 0, 9)$ than when she must choose from C in period 2." (Gul and Pesendorfer 2008, 31).

The authors contrast this representation with the two-parameter model that modifies exponential discounting reviewed in Section 1.3 (e.g. Laibson 1997; O'Donoghue and Rabin 1999, 2001, 2003). Taking the preceding three-period decision problem, the agent's instantaneous utility for each period, u_t, can be represented as follows:

$$u_1(c_1, c_2, c_3) = c_1 + \beta\delta c_2 + \beta\delta^2 c_3$$
$$u_2(c_1, c_2, c_3) = c_2 + \beta\delta c_3$$
$$u_3(c_1, c_2, c_3) = \delta c_3$$

where $\beta > 0$ and $\delta < 1$. Then, according to the discussion above on the normative authority of an *inner or outer rational agent*, it is common practice to take the long-run perspective as the right welfare criterion (as e.g. in O'Donoghue and Rabin 1999, 2003). This corresponds to set $\beta = 1$, which yields the following *fictitious* utility function:

$$u_0(c_1, c_2, c_3) = c_1 + \delta c_2 + \delta^2 c_3$$

which is interpreted as the agent's reconstructed preferences would she not been distorted by a faulty psychologically bias towards the present (e.g. O'Donoghue and Rabin 2003). Gul and Pesendorfer (2008, 31) argue that this welfare criterion is quite arbitrary if one interprets each utility function as a different "self", since, for example, it "assigns a higher welfare to $(1, 0, 11)$ than to $(2, 3, 0)$ even though selves 1 and 2 prefer $(2, 3, 0)$". In my view, this welfare criterion seems somewhat arbitrary because it is not clear why β should be exactly equal to 1 even if one interprets the present bias as a defect.[38] But most importantly, it seems that this welfare criterion applies a *one-size-fits-all* solution that may not be adapted for every person. As argued above, while the behavior that a present bias entails may be indeed considered as a mistake by some persons, it may *not* be by others. This, as argued above, can be conceptualized through the distinction between reflexive and non-reflexive preference change.

[38] O'Donoghue and Rabin (1999, 113-4) argue that from a long-run perspective, even if β is very close to 1 it can create "an arbitrarily large welfare loss" since the rewards and costs can be arbitrarily large or because it is possible to have finitely many periods. It seems to me that this is an insufficient justification for taking $\beta = 1$ as the appropriate welfare criterion for *all* cases (e.g. in a case with few periods and small costs and rewards).

There are two possible cases (considering stable identification preferences). The person does not identify with the present bias, and herself does not want it to be her will. She, and not the observer, considers it to be a defect. In this case it seems sensible to consider a welfare criterion where β is equal or close to 1. If one is sensible to Sugden's (2004) arguments in favor of the sovereignty of first-order preferences and the value of opportunity, one may not endorse any policy that restricts the person's opportunities. However, one may still favor a policy that helps the person to overcome what *she* considers to be a defect. The other possibility is that the agent identifies with the present bias. In this case, the choice to "overrule" a person's preferences (and her choices) is even more controversial than in the standard case without information on second-order preferences. Having the individuals' evaluation of their preferences may then help us to satisfy their sovereignty and the value of opportunity. In addition, either by helping individuals to overcome what they consider to be a defect or abstaining from doing so in cases they do not consider it to be a defect, second-order preferences may provide relevant information to direct resources to people's goals, what they value and seem to care about.

This means that, at least in principle, it is possible to use a multiple preferences approach without constructing a *paternalistic welfare criterion*. By recognizing the person's evolution over time, and the relevance of her evaluation of her preferences, one may respect not only the sovereignty of her preferences but also the sovereignty of her evaluation of those preferences.

1.8 Discussion

In this section I discuss some features, limitations, and potential extensions of the use of second-order preferences in economics. I start with some comments on how to recover second-order preferences (Section 1.8.1), then I briefly discuss the notion of identification (Section 1.8.2), and continue with some of the potential limitations of this framework due to adaptation and false beliefs (Section 1.8.3). I then dis-

cuss some limitations and potential extensions for welfare economics (Section 1.8.4), and finish by distinguishing the view defended here from the inner rational view of agency (Section 1.8.5).

1.8.1 Data

An important feature of any agential theory in economics is how to recover the objects of interest. Traditionally, economics has used choice (or stated choice) as the main source of data. This has led to the identification of the behavioral implications of many theories of choice (often in the form of revealed preference-like axioms). Though in some occasions second-order preferences will translate into choice behavior, other times they will not be reflected on the agents' choices. In this section, I briefly discuss a survey-based and a choice-based method to recover second-order preferences.

Survey-based data, such as individuals' verbal evaluations of their preferences or choices, seems to be one of the most immediate ways of how to recover (evaluative) judgments of an individual. However, as with first-order preferences and choices, second-order preferences may be prone to be affected by frames, cognitive bias, and other sources of context-dependency that limit the reliability of the data collected. If more or less than first-order preferences is an empirical question. But there is sufficient evidence that judgments are prone to bias.[39] For example, Schkade and Kahneman (1998) observed that while students from two Midwest and two California Universities believed that students in California would be significantly happier, the self-reported happiness was very similar in the two locations. This example illustrates a bias/misprediction in judgments concerning adaptation to ways or places of living. They explain this bias through a *focusing illusion*: when reporting their well-being students focused on central aspects of life, while when imagining the happiness of someone else in a different location they focused on the dimensions that differ across

[39]See Kahneman and Thaler (2006) for a review of empirical findings on bias on forecasting/remembering experienced (hedonic) utility.

regions (in this case, climate). They conclude that "[n]othing in life matters quite as much as you think it does while you are thinking about it".

Similar concerns apply to evaluations of past experiences. There is by now some evidence that evaluations of remembered hedonic utility are anchored on the individuals' emotional state when the evaluation takes place (see Stone and Shiffman 1994). For example, high levels of a measure (e.g. pain) on the day of the retrospective evaluation may upwardly bias the retrospective recall of that measure made on that day (Stone, Broderick, Porter and Kaell 1997, 186). If these bias extend to the evaluation of past choices, desires, or preferences is again an empirical question. But this evidence suggests that care should be taken in the design and interpretation of applications of the hierarchical retrospective model.

In addition, survey-based data is often non-incentivized, which may favor inattention and deception. An interesting and controlled way to circumvent some of these limitations is through survey-based experiments. For example, in an experimental setting it is possible to record (in a systematic and controlled way) the duration taken to give an answer in order to exclude *speedy* answers that cannot be the result of honest attentive answers. Another example is the inclusion of incentivized questions of comprehension, which may favor subjects' attention. Though this type of procedures are not perfect solutions, they may increase (if associated with other measures) the reliability of non-choice data.

In a recent survey-based experiment, that brings some evidence that welfare rankings based on (stated) choices are consistent with the happiness view of utility, Benjamin et al. (2012) asked subjects their meta-choices (identification preferences) over binary stated choices in a series of hypothetical choice scenarios. For example, one of their hypothetical choice scenarios was between a job in which the subject would "sleep more but earn less" and a job in which the subject would "sleep less but earn more". Then, for each scenario, they asked subjects the two following questions: "If you were limited to these two options, which do you think you would choose?" (stated choice), followed by "If you were limited to these two op-

tions, which would you want yourself to choose?" (stated meta-choice). Interestingly enough, and bringing some evidence that the conflict between first- and second-order preferences is meaningful, 28% of subjects' stated choices conflicted with their stated meta-choices.[40]

In this experiment there was no time horizon as in the hierarchical preferences models of Section 1.5. But in principle, the elicitation of second-order preferences (either of identification or valuation) can be done either *ex-ante* or in retrospect for most preferences and choices. Note that hypothetical stated choices may sometimes elicit meta-choices, i.e., they may elicit not what the agent's would actually choose when presented with the choice situation but what they would like themselves to choose. Though separating the two questions as in Benjamin et al. (2012) may cue subjects to distinguish between these two notions and give a reasoned answer, this is a problem to have in mind in a survey-based method that pertains to elicit first- and second-order attitudes.

An alternative choice-based method to elicit second-order preferences is with data on *precommitment*, defined as the deliberate restriction of a feasible set (Elster 1982, 222). George (1984, 97) labels these actions as "self-paternalistic", in the sense of an action that "an agent undertakes with the intent of reducing in some way the choice set that he will face at some future time". As the author notes, "the 'revelation' of a meta-preference may be understood to occur via acts of self-paternalism" (p. 95). Though this is certainly not the only behavioral implication of second-order preferences, it seems an important case that can be explained by the conflict between first- and second-order preferences.[41] Choice of precommitment may also be interpreted

[40] In their payed and more controlled experiment run with Cornell University students the percentage increases to 33%. In that experiment some examples of reasons given by subjects for this conflict are: "Sometimes what I want to do may not be what should I do. The choice I should make may be financially better, for instance, but may not be the one I want to choose."; "I made choices I would ideally not want myself to choose in order to secure my future or make friends/family happier. I would probably regret these choices for the reason that life's too short."; "Based on long-term, overall benefits."; "Sometimes I wish my priorities were different than they actually are.".

[41] See George (1984, 96-100) for why other explanations of precommitment, namely (i) that an agent believes he would be "unable" to choose what he *prefers* (instead of what she would want herself to prefer) and (ii) the conflict between an impulsive and a far-sighted self, create difficulties in terms of choice determination and welfare analysis respectively.

as an indication that a second-order preference (e.g. in favor of abstaining from smoking) has more value than a first-order preference (e.g. in favor of smoking) (see also Jeffrey 1974, 383). Combining data on precommitment and retrospective evaluations of choices may be an interesting avenue of research. This could, among other things, bring some evidence on how second-order preferences and self-paternalism are related with regret.

1.8.2 Identification

Throughout this chapter I have used the notion of identification quite broadly, but its precise meaning is somewhat more elusive. Does it mean (precisely) that a person judges a preference/choice as if it is her own in some strong sense? Does it mean that a person judges a preference/choice to reflect who she is or wants to be or become? Or just that a person judges (or even feels) a preference/choice not to be external, refusing sentences of the type "I was not the one to do x". I do not wish to provide here a precise definition, but to briefly revise some of the notions proposed in the literature that are related with second-order attitudes. Among other things, this may help the design of questions for survey-based experiments in order to elicit identification with observed choices.

Though in his early work Frankfurt (1971) seemed to lump together second-order volitions and identification, there is now some agreement that identification consists of higher-order desires/volitions and something more (see Bratman (1996) and Fischer (1999) for reviews). One of the reasons for this is that, as argued above, it is possible that a person does not care and gives no serious consideration to what is at stake when forming a given second-order volition/preference. Another is that one can always refer to a higher-order volition/preference to question if a lower-order volition/preference is really one's own in the strong sense identification seems to require (see e.g. Bratman 2003).[42] Frankfurt himself, in later articles, proposed two

[42] Clearly every choice is one's own in an important (even if) trivial sense. See Watson (1975, 217-9) for the origin of these criticisms.

separate possibilities for the feature that should endorse higher-order attitudes in order to capture identification (and counter these criticisms).

The first was to say that in order to identify with a desire (choice) the second-order attitudes towards this desire (choice) needed to be *wholehearted*. According to Frankfurt (1988), a person acts wholeheartedly when she has "made up her mind" about a decision to take. This requires (roughly) that the decision is *decisive*, in the sense that the person judges that no further consultation of higher-order preferences is needed (see also Frankfurt 1971, 16). The will is undivided and the person is volitionally unified (see also Frankfurt 2009, 91-5). If without reservation or conflict I value and decided to follow my preference to be faithful to my wife, according to this view it seems that no question remains concerning if I value to value (a third-order preference) to be faithful.

Later, the author proposed another criterion based on the *satisfaction* with higher-order attitudes. In his own words, "identification is constituted neatly by an endorsing higher-order desire with which the person is satisfied" (Frankfurt 1992, 14). This requires (again roughly) that a person is settled with respect to the higher-order attitudes over a desire (choice), in the sense that the person *has no interest* in making changes to these higher-order attitudes. It differs from being wholehearted in the sense that it is not an active decision, but just a state in which questions concerning the desirability of the relevant higher-order attitudes do not arise.

An alternative criterion, proposed by Bratman (e.g. 1996), is to consider an endorsement of second-order attitudes based on a decision to treat a desire "as reason-giving in one's practical reasoning and planning concerning some relevant circumstances" (p. 9). According to the author, "[t]o identify with one's desire is (a) to reach a decision to treat that desire as reason-giving and to be satisfied with that decision, and (b_1) to treat that desire as reason-giving or, at least, (b_2) to be fully prepared to treat it as reason-giving were a relevant occasion to arise" (Bratman 1996, 12). Though all these criteria seem to have their own shortcomings, the discussion around identification suggests that the following is true:

> "It appears, then, that the special status of higher-order volitions must be explained by something other than the fact that they are desires of a higher order. It must be explained by the fact that they are endorsed by the agent in acts of identification or decisive commitments." (Lippert-Rasmussen 2003, 354; see also Watson 1975, 217-9)

1.8.3 Adaptive and Informational Preference Change

It is important to distinguish preference change due to the resolution of a conflict between a first- and second-order preference (reflexive preference change), from other phenomena that also tend to change preferences over time. I will briefly discuss two of these phenomena here: (i) adaptive preference change and (ii) informational preference change. In what follows, I draw upon Elster (1982) that is a classical (and in many ways comprehensive) treatment of adaptive preferences, and Cowen (1993) and Harsanyi (1997) for treatments of the question of preference change due to learning and/or experience.

Adaptive preference change, according to Elster (1982), refers to the cases in which aspirations are downgraded due to restrictions in the feasible set.[43] For example, say someone prefers job x that she can get if promoted to her current job y. Before the decision of promotion takes place, her feasible set of possible outcomes (at least in terms of her aspirations) is $\{x, y\}$. But if she does not get the promotion, her feasible set (in this sense) becomes $\{y\}$. Then, if that is the case, she may rationalize the non-promotion by saying that "the top job is not worth having anyway", changing her preferences for y over x (see Elster 1982, 225).

As exposed by Elster (1982), adaptive preferences (or preference change) differs in important ways from reflexive preferences (or change). The first distinction is

[43] This contrasts with a preference for what one does not have (e.g. I prefer to be single to be married when I am married, but I prefer to be married to be single when I am single). Elster (1982, 226) also distinguishes adaptive preference change "from learning in that it is reversible; from precommitment in that it is an effect and not a cause of a restricted feasible set; from manipulation in that it is endogenous; from character planning [reflexive preference change] in that it is causal; and from wishful thinking in that it concerns the evaluation rather than the perception of the situation".

that adaptive preference formation or change takes place "behind the back" of the person, as a causal (non-conscious) drive, while reflexive preference change (or deliberate character planning in Elster 1982, 224) is in general deliberate, conscious, and intentional.

Another distinction, according to Elster (1982, 237-8), is that in order to treat adaptive preferences one needs to consider the "*genesis* of wants", which involves "an inquiry into the history of the actual preferences". This means looking at the sequence of choices as revealing information about the *nature* of these choices. Here, I have taken the nature of choices to be mostly revealed through stated second-order attitudes (at least when one is interested in distinguishing between reflexive and non-reflexive choices). One could instead use the sequence of first- and second-order preferences in order to try to capture some patterns that give us indications if preferences have been reflexive, adaptive, or other. The potential of such line of inquiry is mostly an open question.

A third and important distinction, according to Elster (1982, 235), is that while reflexive preference change may "improve welfare without loss of autonomy", that is not the case of adaptive preference change. Even though adaptive preferences may improve welfare (e.g. due to resignation and reduction of frustration), they do so in a non-autonomous way.[44] Elster (1982, 235) recognizes that second-order attitudes may not be autonomous, but argues that such cases are not centrally important. I think that such cases should be taken seriously, since, as argued in Section 1.8.1, evaluative judgments are also prone to adaptation and other bias.

Informational preference change refers to the cases in which preferences change due to learning and/or experience. For example, a patient may prefer treatment x over y based on false beliefs about the secondary effects of treatment x. Informed about these effects, she may change her preferences for y over x. This is different from reflexive preference change since it does not *necessarily* relate with identifica-

[44]Though Elster (1982) does not provide a definition or criterion for autonomous wants, he consider adaptation (in the sense defined above) as a mechanism that shapes individual wants in a non-autonomous way (see p. 228).

tion, nor identification necessarily relates with new information, though new information may be one way in which one changes her identification over some preferences/choices. This kind of examples led many authors to endorse different versions of fully-informed preferences for policy and welfare evaluations.[45] It is then useful to distinguish fully-informed preferences from reflexive preferences.

There are at least two ways of defining fully-informed preferences. I call them the *hypothetical informed preferences* and *actual informed preferences*. The former are the *hypothetical* preferences the agent would have if she had all the relevant information and made full use of this information (e.g. Harsanyi 1997, 133). Sometimes this version of informed preferences is coupled with some demands of rationality in terms of the use of information. The latter are the individual preferences the agent would *actually* have if she would be informed of all the relevant information. This version therefore involves no demand of full or rational use of information. Actual informed preferences are the agent's own preferences after being informed of all the relevant information. This relevant information may include info on cognitive bias that often affect agents' preferences, such as framing or adaptation.

Informed preferences face some difficulties of their own. For one, it is not necessarily straightforward to define what is the *relevant* information. How much and what type of information is necessary for one to be fully-informed about some topic? Second, informed preferences, either actual or hypothetical, are very often not available. They do not exist *now* as they pertain to the preferences with information (and cognitive capacities in the hypothetical case) that the agent does not have:

> "The preferences of perfectly informed individuals are not always relevant for imperfectly informed choice. By considering perfectly informed preferences, we are hypothetically changing an individual's human capital endowment. What an individual would want with a different human capital endowment cannot necessarily be extrapolated usefully into infor-

[45] From which Hausman (2012) is a notable example. See Cowen (1993) for an old but interesting review.

> mation about what improves the welfare of an individual now" (Cowen 1993, 262)

Finally, informed preferences, in particular hypothetical ones, may be against individual sovereignty. Contrary to reflexive preferences, informed preferences refer to preferences that are not of the individuals themselves. For example, Harsanyi (1997) defines "mistaken preferences" as actual preferences that are based on some objective view of incorrect or incomplete information. Then, one of his proposals is to base the welfare evaluation of one individual on the "preferences of *other* knowledgeable people" (see Harsanyi 1997, 134 and 142-3). But one can certainly find many examples where the preferences of most knowledgeable people do not respect one's preferences.[46]

Despite these limitations, informed preferences have the potential to "purify" preferences of cases of blatant false beliefs, as the example of medical treatments highlights. This is something that, as it was the case with adaptation, escapes from the notion of reflexive preferences. Then, a full-fledged theory of preferences and well-being may need some kind of adaptation and informational criteria if they are to be used for policy and welfare analysis.

One standard strategy, as suggested by Harsanyi (1997), is to define some actions as *objective mistakes*, i.e., to judge some actions as against the agent's own interest even if the agent does not necessarily agree. A similar strategy is to use an observer's meta-ranking. According to Amartya Sen (1974, 1977, 1980), instead of simply looking to the persons' multiple preferences over alternatives we should also rank their multiple preferences according to some social desirable criteria. If preferences based on adaptation and false beliefs are judged negatively for a given context, they could be low ranked in a partial or complete ordering of different preferences according to their social or moral worth. According to Sen (1977, 338), this kind of meta-rankings would endow an observer with "a varying extent of moral articulation". An observer's

[46] Harsanyi (1997, see 134) defends a *negative* version of paternalism, in which *we* do not coerce another into what we think is the correct behavior but we only refuse to subsidize activities that we consider to be against the agent's own interest.

ranking over the multiple preferences could, if information and confidence on social judgments are sufficient, rank in a social or moral convenient way preferences based on false beliefs and adaptation, as well as preferences based on expensive tastes, antisocial desires, coercion or manipulation.

But these strategies are at the cost of individual sovereignty. I have tried to design a model where mistakes are instead self-authenticated. This could, at least in principle, be endorsed both by author in favor or against paternalism. I have also defended that judgments, much like choices, are important if we want to respect a broad (positive) view of individual sovereignty. So how to to save this model in face of uninformed and adaptive preferences?

One circuitous strategy would be to define the adaptation and informational criteria with respect to the *context* of interest, i.e., to define a minimum degree of information and non-adaptation in order to use first- and second-order preferences as guidance for positive and normative analysis. This would allow to keep a hierarchical structure without defining objective mistakes with respect to observed choices as long as the context is "information and adaptation proof". But of course, this is silent with respect to the other contexts where lack of crucial information and adaptation seem to be important.

Another possibility for those, as myself, that care about individual sovereignty but think that false beliefs and adaptation are important and real issues would be to respect first- and second-order preferences at the same time as implementing "experimentation" (non-coercive and non-obligatory) policies aimed to deal with these issues. Elster (1982, 221) gives an example to deal with lack of information with respect to different ways of living: "a systematic policy of experimentation that gave individuals an opportunity to learn about new alternatives without definite commitment". Of course, in an economy with limited resources, it may be that this kind of policies may be only implemented at the expense of some others that favor the actual preferences and judgments of individuals (implying a kind of negative paternalism; see footnote 46). This kind of policy could also procure a loss in terms of the surprise

and discovery that certain goods procure to an individual (see Cowen 1993, 263). What to do is a hard question that I wish to leave open.

1.8.4 The Objective and Intrinsic Values of Opportunity

As suggested in the previous sections, the satisfaction of preferences may not be a good indicator of individual and social welfare for a number of reasons besides the conflict between first- and second-order preferences. Adaptation, false beliefs, or antisocial preferences are just a few examples of the difficulties with this stand (see e.g. Hausman 2012, 81-2). In addition, two important dimensions that a normative framework based on information on first- and second-order preferences would fall short are: (i) the *objective value* of opportunity, and (ii) the *intrinsic value* of opportunity. With respect to the first, by relying exclusively on first- and second-order preferences it is possible to have welfare criteria compatible with deficient opportunities, individual rights, or basic capabilities. This is an uncontroversial deficiency of a normative framework for the ones, as myself, that believe that welfare criteria based on preferences should be accompanied by minimal and distributional considerations of objective measures of well-being such as real opportunities or capabilities.

With respect to the second, the case is that the freedom or the opportunities one has may be valued *for themselves*. For instance, one may prefer a leader to be selected through an election rather than through appointment, irrespective of the identity of the leader that is finally chosen. Freedom in this sense is an end in itself. Neither first- nor second-order preferences over alternatives take this intrinsic value of opportunities into account. A potential escape to this limitation would be to follow Gravel (1994, 1998), who defines first-order preferences on the set of alternatives and on the opportunity sets that contain them, such that preferences are defined over some *extended* set of ordered pairs like (a, A) or $(a, \{a\})$. This allows to express statements such as "I prefer choosing a from set A to choosing the same a from set $\{a\}$" (Gravel 1994, 455). Second-order preference would be then defined over a

pair of preferences, indicating if an individual identifies or not with her preferences over alternatives *and* her preferences over opportunity sets. An open question is if using second-order preferences would be useful to overcome some of the difficulties identified by Gravel (1994, 1998) with finding an ordering of the extended set of ordered pairs that respects individual preferences.

1.8.5 Hierarchical Models and the Inner Rational Agent

A question that may emerge from the reading of this chapter is if adopting hierarchical models is not pushing the inner rational agent model one level up. In fact, when meta-preferences are discussed in economics they are often (implicitly or explicitly) assumed to be a stable and context-independent single ranking of multiple preferences. This corresponds, in some sense, to an inner rational agent model at the second-order level. The same is true if one assumes the existence of consistent *latent* second-order preferences which can be reconstructed by eliminating objective mistakes.

I would like to argue that the view exposed in this chapter, in particular in the two hierarchical preferences models, is not necessarily in line with the inner rational view of agency. In the retrospective model, second-order preferences (both identification and valuation) are not assumed to be neither transitive nor complete (though they are assumed to be acyclic), and most importantly, they are assumed to be the result of a reflexive activity in *one period*. This means that the model does not impose stability of second-order preferences. I have neither assumed that second-order preferences are context-independent. In the evolving model, the agent is assumed to have the reflexive ability to evaluate his preferences/choices in several instances, but second-order preferences are again not assumed to be rational in terms of transitivity, completeness, and most importantly, stability or context-independence. Finally, all preferences (choices) are self-authenticated and second-order preferences cannot

be judged mistaken from the *outside* within these models. This seems to contrast with most versions of the inner (and outer) rational agents found in the literature.

This of course is at some cost. The limits in terms of behavioral implications and welfare analysis of a model in which preferences and higher-order attitudes may be context-dependent and/or evolve over time is an open (and in many ways empirical) question. But one would certainly lose some parsimony and power in terms of prediction of behavior. In terms of welfare analysis, we have seen that more structure is needed in order to *guarantee* an acyclic welfare ranking that takes choices and reflexive preferences into account. Similarly, context-dependence upon the order at which stated identification preferences are elicited, could pose a problem for the use of such information in welfare analysis. In sum, I believe that many challenges face someone who wishes to take the time and context-dependency of preferences and higher-order attitudes seriously into consideration. But some strategies may be available, such as finding "adaptation and information proof" contexts to elicit first- and second-order preferences. In such settings, one could be more confident on fully respecting the preferences and judgments of individuals and on finding coherent and/or acyclic welfare rankings even though acyclicity would not be theoretically guaranteed.

1.9 Concluding Remarks

From this analysis, it seems that economics could gain from adopting a richer conception of the economic agent that accounts for some of the essential capacities that define a *person*, such as the ability to evaluate and change one's preferences, one's personal identity, or *who* one is. A person, according to the view exposed here, can either identify or not with her preferences and her preference change. Among other advantages, considering a *person instead of an agent* could get economics closer to what people care about, their goals, who they want to be or become, and what is important to them.

One can contrast this view with Sugden's (2004, 2007) proposal to model the economic agent as a "continuing agent", that can be seen "as the *composition* of the series of *time-slice agents*" (i.e., the composition of the agent in period 1, the agent in period 2, and so on) [see 2007, 671]. According to this view, an agent's choice in period t is not only interpreted as the deliberate choice of the "agent in period t" but also as the deliberate choice of the "continuing agent". Sugden (2004, 2007) sees this continuing agent as a continuing *locus of responsibility* and argues that this agent "*identifies* with each of his time-slices" (see 2007, footnote 5). What my arguments suggest is that there are many instances when people do not identify with what they have done or are about to do, and that this provides meaningful information to understand the nature of individuals' decisions.

Several questions are left open regarding how these notions relate with well-being, justice, and moral responsibility. Take the example of Jack, a husband that has betrayed his wife although he would have liked to remain faithful. According to his second-order preferences, Jack seems not to identify with his time-slice that betrayed his wife. Is this an indication that he regrets his (free) choice, and that this action has decreased his hedonic well-being? Is it fair, say for questions of "punishments and/or rewards", to differentiate a case like Jack's from that of a husband that identifies with his betrayal? Should the fact that Jack does not identify with his action excuse him of any or some moral responsibility? These questions illustrate the fascinating topics that are left to explore.

Finally, I have taken a deterministic view of agency but non-deterministic views of agency such as *random preferences models*, according to which agents' preferences change stochastically (e.g. Becker, DeGroot and Marschak 1963; Barberà and Pattanaik 1986; McFadden and Richter 1990; Loomes and Sugden 1995; Gul and Pesendorfer 2006; Apesteguia, Ballester and Lu 2017), or *deliberate randomization models*, according to which agents deliberately choose stochastically following a preference to reduce regret, incomplete preferences, difficulty to judge one's true risk aversion, and the like (e.g. Machina 1985; Marley 1997; Fudenberg and Strzalecki

2014), also rationalize (and predict) preference change.[47] These models could explain and rationalize, for example, that *ceteris paribus* an individual is more likely to change behavior when her first and second-order preferences conflict than when they don't. They are also alternative deliberate rationalizations of choices that are often judged in economics as mistakes. In fact, in a recent experiment with repeated choices on similar lotteries, Agranov and Ortoleva (2017) disentangle if stochastic choice (i.e., different choices when choosing from the same set of alternatives many times) can be rationalized by these models or by subjects' mistakes. Their main finding is that the majority of subjects choose stochastically on purpose rather than commit mistakes. This provides yet another argument in favor of not interpreting, *a priori*, any deviation of "rational" behavior as a mistake.

[47] See Fishburn (1999) for an old but comprehensive review.

Bibliography

Afriat, S. (1967) The Construction of Utility Functions from Expenditure Data. International Economic Review 8(1): 67–77.

Agranov, Marina and Pietro Ortoleva (2017) Stochastic Choice and Preferences for Randomization. Journal of Political Economy 125(1): 40–68.

Aizerman, M. A. and A. V. Malishevski (1981) General Theory of Best Variants Choice: Some Aspects. IEEE Transactions on Automatic Control 26(5): 1030–41.

Akerlof, G. A. (1991) Procrastination and Obedience. The American Economic Review 81(2): 1–19.

Apesteguia, J. and M. A. Ballester (2010) The Computational Complexity of Rationalizing Behavior. Journal of Mathematical Economics 46: 356–63.

——— (2015) A Measure of Rationality and Welfare. Journal of Political Economy 123(6): 1278–1310.

Apesteguia, J., M. A. Ballester, and J. Lu (2017) Single Crossing Random Utility Models. Econometrica 85(2): 661–74.

Arlegi, R. and M. Teschl (2015) Conflicts in Decision Making. In: C. Binder, G. Codognato, M. Teschl, and Y. Xu (eds) Individual and Collective Choice and Social Welfare: Essays in Honor of Nick Baigent. Springer-Verlag Berlin Heidelberg, : 11–29.

Arrow, K. J. (1951) Social Choice and Individual Values. Wiley, New York.

——— (1959) Rational Choice Functions and Orderings. Economica 26: 121–27.

Baigent, N. (1995) Behind the Veil of Preferences. Japanese Economic Review 46(1): 88–101.

Barberà, S. and P. K. Pattanaik (1986) Falmagne and the Rationalizability of Stochastic Choices in Terms of Random Orderings. Econometrica 54(3): 707–15.

Becker, G., M. DeGroot, and J. Marschak (1963) Stochastic Models of Choice Behavior. Behavioral Science 8: 41–55.

Benjamin, D. J., J. J. Choi, and A. J. Strickland (2010) Social Identity and Preferences. The American Economic Review 100(4): 1913–28.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and A. Rees-Jones (2012) What Do You Think Would Make You Happier? What Do You Think You Would Choose?. The American Economic Review 102(5): 2083–110.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and N. Szembrot (2014) Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference. The American Economic Review 104(9): 2698–735.

Berlin, I. (1969) Four Essays on Liberty. Oxford University Press., Oxford.

Bernheim, B. D. and A. Rangel (2004) Addiction and Cue-Triggered Decision Processes. The American Economic Review 94(5): 1558–90.

——— (2007) Toward Choice-Theoretic Foundations for Behavioral Welfare Economics. The American Economic Review: Papers and Proceedings 97(2): 464–70.

——— (2009) Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. The Quarterly Journal of Economics 124(1): 51–104.

Bossert, W., Y. Sprumont, and K. Suzumura (2005) Consistent Rationalizability. Economica 72: 185–200.

——— (2006) Rationalizability of Choice Functions on General Domains Without Full Transitivity. Social Choice and Welfare 27: 435–58.

Bowles, S. (1998) Endogenous Preferences: The Cultural Consequences of Markets and other Economic Institutions. Journal of Economic Literature 36: 75–111.

Bratman, M. E. (1996) Identification, Decision, and Treating as a Reason. Philosophical Topics 24(2): 1–18.

——— (2003) A Desire of One's Own. The Journal of Philosophy 100(5): 221–42.

Brennan, T. J. (1993) The Futility of Multiple Utility. Economics and Philosophy 9: 155–64.

Broome, J. (1991) Weighing Goods. Basil Blackwell, Oxford.

Bykvist, K. (2003) The Moral Relevance of Past Preferences. In: H. Dyke (ed) Time and Ethics: Essays at the Intersection. Kluwer, Dordrecht, Holland: 115–36.

Carpenter, J. P. (2005) Endogenous Social Preferences. Review of Radical Political Economics 37(1): 63–84.

Cowen, T. (1993) The Scope and Limits of Preference Sovereignty. Economics and Philosophy 9: 253–69.

Davis, J. B. (2009) Identity and Individual Economic Agents: A Narrative Approach. Review of Social Economy 67(1): 71–94.

Decancq, K., M. Fleurbaey, and E. Schokkaert (2015) Happiness, Equivalent Incomes and Respect for Individual Preferences. Economica 82: 1082–106.

Dennett, D. (1991) Consciousness Explained. Little Brown & Co.

Dietrich, F. and C. List (2013) Where do Preferences Come From?. International Journal of Game Theory 42(3): 613–37.

——— (2016) Reason-based choice and context-dependence: An explanatory framework. Economics and Philosophy 32(2): 175–229.

Elster, J. (1982) Sour Grapes - Utilitarianism and the Genesis of Wants. In: A. K. Sen and B. Williams (eds) Utilitarianism and Beyond. Cambridge University Press, Cambridge: 219–38.

——— (1985) Introduction. In: J. Elster (ed) The Multiple Self. Cambridge University Press, Cambridge: 1–34.

Fehr, E. and K. Hoff (2011) Introduction: Tastes, Castes and Culture: The Influence of Society on Preferences. The Economic Journal 211: 396–412.

Fischer, J. M. (1999) Recent Work on Moral Responsibility. Ethics 110(1): 93–139.

Fishburn, P. (1999) Stochastic Utility. In: S. Barberà, P. Hammond, and C. Seidl (eds) Handbook of Utility Theory. Vol. 1 Principles Kluwer Academic Publishers, Dordrecht, Holland: 273–320.

Fleurbaey, M. and E. Schokkaert (2013) Behavioral Welfare Economics and Redistribution. American Economic Journal: Microeconomics 5(3): 180–205.

Foster, J. (1993) Notes on Effective Freedom. Mimeo, Vanderbilt University.

Frankfurt, H. G. (1971) Freedom of the Will and the Concept of a Person. The Journal of Philosophy 68(1): 5–20.

——— (1988) Identification and Wholeheartedness. In: The Importance of What We Care About: Philosophical Essays. Cambridge University Press.

——— (1992) The Faintest Passion. In: Proceedings and Addresses of the American Philosophical Association. 66 American Philosophical Association., Newark, Del.: 5–16.

——— (2009) The Reasons of Love. Princeton University Press, Princeton, N.J.

Fudenberg, D. and D. K. Levine (2006) A Dual Self Model of Impulse Control. The American Economic Review 96: 1449–76.

——— (2012) Timing and Self-Control. Econometrica 80(1): 1–42.

Fudenberg, D. and T. Strzalecki (2014) Recursive Stochastic Choice. , Mimeo, Harvard University.

Gallagher, S. (2000) Philosophical Conceptions of the Self: Implications for Cognitive Science. Trends in Cognitive Sciences 4(1): 14–21.

George, D. (1984) Meta-Preferences: Reconsidering Contemporary Notions of Free Choice. International Journal of Social Economics 11(3/4): 92–107.

Gravel, N. (1994) Can a Ranking of Opportunity Sets Attach an Intrinsic Importance to Freedom of Choice?. American Economic Review Papers and Proceedings 84: 454–58.

——— (1998) Ranking Opportunity Sets on the Basis of their Freedom of Choice and their Ability to Satisfy Preferences: A Difficulty. Social Choice and Welfare 15: 371–82.

——— (2008) What is Freedom?. In: Handbook of Economics and Ethics. Edward Edgar Publishing, London.

Gul, F. and W. Pesendorfer (2001) Temptation and Self-control. Econometrica 69(6): 1403–36.

——— (2004) Self-control and the Theory of Consumption. Econometrica 72(1): 119–58.

——— (2005) The Revealed Preference Theory of Changing Tastes. The Review of Economic Studies 72(2): 429–48.

——— (2006) Random Expected Utility. Econometrica 74(1): 121–46.

——— (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: 3–39.

Harsanyi, J. C. (1997) Utilities, Preferences, and Substantive Goods. Social Choice and Welfare 14: 129–45.

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

——— (2013) A reply to Lehtinen, Teschl and Pattanaik. Journal of Economic Methodology 20(2): 219–23.

Hirschman, A. O. (1984) Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse. The American Economic Review: Papers and Proceedings 74(2): 89–96.

Hoff, K. and J. E. Stiglitz (2016) Striving for Balance in Economics: Towards a Theory of the Social Determination of Behavior. Journal of Economic Behavior and Organization 126: 25–57.

Horst, U., A. Kirman, and M. Teschl (2006) Changing Identity: The Emergence of Social Groups. GREQAM Working Paper.

Infante, G., G. Lecouteux, and R. Sugden (2016) Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics. Journal of Economic Methodology 23(1): 1–25.

Jamison, J. and J. Wegener (2010) Multiple Selves in Intertemporal Choice. Journal of Economic Psychology 31: 832–39.

Jeffrey, R. C. (1974) Preferences Among Preferences. The Journal of Philosophy 71(13): 377–91.

Jones, P. and R. Sugden (1982) Evaluating Choice. International Review of Law and Economics 2: 47–65.

Kahneman, D. (2003) Maps of Bounded Rationality: Psychology for Behavioral Economics. The American Economic Review 93(5): 1149–75.

——— (2011) Thinking, Fast and Slow. Farrar, Straus & Giroux, New York, NY.

Kahneman, D. and R. H. Thaler (2006) Anomalies: Utility Maximization and Experienced Utility. The Journal of Economic Perspectives 20(1): 221–34.

Kalai, G., A. Rubinstein, and R. Spiegler (2002) Rationalizing Choice Function by Multiple Rationales. Econometrica 70(6): 2481–88.

Kirman, A. and M. Teschl (2006) Searching for Identity in the Capability Space. Journal of Economic Methodology 13(3): 299–325.

Koszegi, B. and M. Rabin (2007) Mistakes in Choice-based Welfare Analysis. American Economic Review Papers and Proceedings 97(2): 477–81.

Laibson, D. (1994) Essays in Hyperbolic Discounting. Ph.D. dissertation, MIT.

——— (1997) Golden Eggs and Hyperbolic Discounting. The Quarterly Journal of Economics 112(2): 443–77.

LeBoeuf, R. A., E. Shafir, and J. B. Bayuk (2010) The Conflicting Choices of Alternating Selves. Organizational Behavior and Human Decision Processes 111(1): 48–61.

Lehtinen, A. (2012) A Review on Daniel Hausman (2012): Preference, Value, Choice, and Welfare.

Lewis, D. (1989) Dispositional Theories of Value. Proceedings of the Aristotelian Society, Supplementary Volumes 63: 113–137.

Lippert-Rasmussen, K. (2003) Identification and Responsability. Ethical Theory and Moral Practice 6: 349–76.

Livet, P. (2006) Identities, Capabilities and Revisions. Journal of Economic Methodology 13(3): 327–48.

Loomes, G. and R. Sugden (1995) Incorporating a Stochastic Element Into Decision Theories. European Economic Review 39: 641–48.

Machina, M. J. (1985) Stochastic Choice Functions Generated from Deterministic Preferences over Lotteries. The Economic Journal 95: 575–94.

Manzini, P. and M. Mariotti (2007) Sequentially Rationalizable Choice. The American Economic Review 97(5): 1824–39.

Marley, A. (1997) Probabilistic Choice as a Consequence of Nonlinear (sub) Optimization. Journal of Mathematical Psychology 41: 382–91.

McFadden, D. and M. K. Richter (1990) Stochastic Rationality and Revealed Stochastic Preference. In: J. S. Chipman, D. McFadden, and M. K. Richter (eds) Preferences, Uncertainty, and Optimality: Essays in Honor of Leo Hurwicz. Westview Press: Boulder, CO, 161-186., Boulder, Colorado: 163–186.

Nehring, K. and C. Puppe (1999) On the Multi-preference Approach to Evaluating Opportunities. Social Choice and Welfare 16: 41–63.

O'Donoghue, T. and M. Rabin (1999) Doing It Now or Later. The American Economic Review 89(1): 103–24.

——— (2001) Choice and Procrastination. The Quarterly Journal of Economics 116(1): 121–60.

——— (2003) Studying Optimal Paternalism, Illustrated with a Model of Sin Taxes. The American Economic Review: Papers and Proceedings 93(2): 186–91.

Parfit, D. (1984) Reasons and Persons. Oxford University Press, Oxford.

Pattanaik, P. K. and Y. Xu (1990) On Ranking Opportunity Sets in Terms of Freedom of Choice. Recherches Economiques de Louvain 56: 383–90.

Rabin, M. (2013) Incorporating Limited Rationality into Economics. Journal of Economic Literature 51(2): 528–43.

Ricoeur, P. (1984) Time and Narrative (3 Vols). University of Chicago Press.

——— (2002) Narrative Identity. In: D. Wood (ed) On Paul Ricoeur: Narrative and Interpretation. Routledge, London.

Rubinstein, A. and Y. Salant (2012) Eliciting Welfare Preferences from Behavioural Data Sets. The Review of Economic Studies 79: 375–87.

Schelling, T. (1984) Choice and Consequence: Perspectives of an Errant Economist. Harvard University Press, Cambridge, MA.

Schkade, D. A. and D. Kahneman (1998) Does Living in California Make People Happy? A Focusing Illusion in Judgments of Life Satisfaction. Psychological Science 9(5): 340–46.

Schotter, A. (2008) What's so Informative About Choice?. In: A. Caplin and A. Schotter (eds) The Foundations of Positive and Normative Economics. Oxford University Press, New York: 70–94.

Sen, A. K. (1971) Choice Functions and Revealed Preferences. Review of Economic Studies 38: 307–17.

——— (1974) Choice, Orderings, and Morality. In: S. Koerner (ed) Practical Reason, Oxford: Basil Blackwell.. Basil Blackwell, Oxford.

——— (1977) Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. Philosophy and Public Affairs 6(4): 317–44.

——— (1980) Plural Utility. In: Proceedings of the Aristotelian Society, New Series. 81 Wiley on behalf of The Aristotelian Society, : 193–215.

——— (1987) Commodities and Capabilities. Oxford University Press, New Delhi, Indiaoxford india paperbacks 1999 edition.

——— (1988) Freedom of Choice: Concept and Content. European Economic Review 32: 269–94.

——— (1991) Welfare, Preference and Freedom. Journal of Econometrics 50: 15–29.

Stigler, G. J. and G. S. Becker (1977) De Gustibus Non Est Disputandum. The American Economic Review 67(2): 76–90.

Stone, A. A., J. E. Broderick, L. S. Porter, and A. T. Kaell (1997) The Experience of Rheumatoid Arthritis Pain and Fatigue: Examining Momentary Reports and Correlates over one Week. Arthrities & Rheumatology 10(3): 185–93.

Stone, A. A. and S. Shiffman (1994) Ecological Momentary Assessment (EMA) in Behavioral Medicine. Annals of Behavioral Medicine 16: 199–202.

Strotz, R. H. (1955) Myopia and Inconsistency in Dynamic Utility Maximization. Review of Economic Studies 23(3): 165–80.

Sugden, R. (2004) The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences. The American Economic Review 94(4): 1014–33.

——— (2007) The Value of Opportunities over Time when Preferences are Unstable. Social Choice and Welfare 29: 665–82.

Thaler, R. H. and H. M Shefrin (1981) An Economic Theory of Self Control. Journal of Political Economy 89(2): 392–406.

Varian, H. (2006) Revealed Preferences. In: M. Szenberg L. Ramrattan and A. Gottesman (eds) Samuelsonian Economics and the Twenty First Century. Oxford University Press, Oxford: 99–115.

Watson, G. (1975) Free Agency. The Journal of Philosophy 72: 205–20.

Chapter 2

The Tree that Hides the Forest: A Note on Revealed Preference

The common interpretation given to choice behavior that satisfies the traditional revealed preference axioms is that it results from the maximization of a single preference. I show that choice data alone does not enable one to rule out the possibility that the choice behavior that satisfies the revealed preference axioms is instead the result of the aggregation of a collection of distinct preferences. In particular, I show that any ordering is observationally equivalent to a majoritarian aggregation of a collection of distinct dichotomous orderings. I also show that any ordering is observationally equivalent to a Borda's aggregation of a collection of distinct linear orderings. I use these two examples and other related results to discuss the topics of (in)distinguishability and model selection.

Keywords: Revealed preference theory; Rationalization; Dichotomous preferences; Aggregation rules; Distinguishability; Model selection.

> "It happens that C [Weak and Strong Axioms of Revealed Preference] implies B [maximizing ordinal utility] as well as being implied by it. It is nonsense to think that C could be realistic and B unrealistic, and nonsense to think that the unrealism of B could then arise and be irrelevant." (Samuelson in Archibald, Simon and Samuelson 1963, 235)

> "In the standard approach, the terms "utility maximization" and "choice" are synonymous. A utility function is always an ordinal index that describes how the individual ranks various outcomes and how he behaves (chooses) given his constraints (available options). The relevant data are revealed preference data, that is, consumption choices given the individuals constraints." (Gul and Pesendorfer 2008, 7)

2.1 Introduction

Initiated by Samuelson (1938) and developed by Houthakker (1950), Arrow (1959), Richter (1966), Afriat (1967), Sen (1971) among many others, the revealed preference theory has established an equivalence between observable properties of choice behavior - the traditional revealed preference axioms - and the possibility for this behavior to be rationalized by a single preference.[1] When choice behavior satisfies the revealed preference axioms, it is then common to interpret it, as the initial quotations illustrate, as an indication that this behavior results from the maximization of a single preference.

In this chapter, I show that it is not possible to distinguish - using choice data alone - if the behavior of an agent (be it an individual or an institution) that satisfies the traditional revealed preference axioms is the result of a (direct) maximization of a single preference *or* the result of the aggregation of a collection of distinct preferences. I show this for two widely known and used aggregation rules. First, I show that *any* ordering is (observationally) equivalent to an aggregation of a collection of distinct dichotomous orderings by the majority rule. Second, I show that *any* ordering is (observationally) equivalent to an aggregation of a collection of distinct linear orderings by the Borda's rule.

[1] Examples of these axioms include the Weak and Strong Axioms of Revealed Preference and the Weak and Strong Congruence Axioms. As shown by Sen (1971), these axioms are all equivalent when applied to a choice correspondence defined over a finite set that includes all two-element and three-element sets.

The interest of these results is not that other interpretations whatsoever are possible. As it is well known in the literature, the satisfaction of the revealed preference axioms is just an indication that the observable behavior can be described *as if* it is the result of the maximization of a single preference (as opposed to the confirmation that this hypothesis is true). The interest is instead that some commonly used *collective* decision rules (or multiple preferences models) have the same empirical implications in terms of choice than the standard *individual* rational choice model.

Take the example of an organization (e.g. a firm, government, or household) that makes a series of choices consistent with the revealed preference axioms. The results of this chapter imply that this behavior may not be the result of a "dictatorial" decision, but instead the result of a collective decision (the aggregation of the preferences of the several members of the organization). Alternatively, suppose that one observes the choices of an individual decision maker. These results indicate that if his or her choices satisfy the revealed preference axioms, then there exists an internal process by which a collection of individual "selves" are aggregated into a single ordinal index.

I regard these results as a way to unveil several interesting questions, rather than a suggestion that the impossibility to distinguish two models by means of revealed preference data is by itself problematic. How plausible are these models as explanations of some behavior? When is it important to know the underlying model of a given choice behavior? With answers to these two questions, one may then want to ask: What to do in order to distinguish between two or more plausible models that are not distinguishable through choice data?

The remainder of the chapter is organized as follows. Section 2.2 is devoted to notation and preliminaries. In Section 5.5 I present the results. In Section 2.4 I discuss the questions just posed. I summarize the contribution of this chapter in Section 2.5.

2.2 Notation and Preliminaries

Let X be a finite set of alternatives denoted by x, y, etc., and assume that $\#X = n \geq 3$. Let A be any subset of X and $\mathcal{P}(X)$ denote the set of all non-empty subsets of X. A choice correspondence is a mapping $C : \mathcal{P}(X) \to \mathcal{P}(X)$ that satisfies $C(A) \in A$ for every $A \in \mathcal{P}(X)$. The standard interpretation is that $C(A)$ is the choice set (or consideration set) from the set A. The choice set must contain at least one alternative, and if it has more than one they are interpreted as equally good *potential* choices.[2]

Let $I^1, ..., I^{m(\succsim)}$ denote the $m(\succsim)$ *indifference classes* of $\succsim$ from top to bottom such that for $1 \leq j < l \leq m(\succsim)$ if $x, y \in I^j$ then $x \sim y$ and if $x \in I^j$ and $y \in I^l$ then $x \succ y$. An ordering $\succsim$, i.e., a reflexive, complete, and transitive binary relation on X, is said to be *dichotomous* if $m(\succsim) \leq 2$.

For any ordering $\succsim$, a choice correspondence C is said to be *rationalized* by $\succsim$ if $C(A) = \{x \in A : \ x \succsim y \ \forall y \in A\}$ for all $A \in \mathcal{P}(X)$. One of the traditional revealed preference axioms, the Weak Axiom of Revealed Preference (WARP), states that if an alternative x is "revealed preferred" to y (i.e., x is once chosen when y is available *and* rejected), then y is *not* revealed to be "at least as good as" x (i.e., y is never chosen when x is available). Formally, WARP states that for all $x, y \in X$ and $A, B \in \mathcal{P}(X)$ if $x \in C(A)$ and $y \in A \setminus \{C(A)\}$ then $\neg(y \in C(B)$ and $x \in B)$. A choice correspondence that satisfies this condition can be viewed as resulting from the maximization of a single preference, i.e.:

Theorem 1. *(Arrow 1959) A choice correspondence C satisfies WARP if and only if it is rationalized by an ordering $\succsim$.*

For any collection $(\succsim_1, ..., \succsim_k)$ of k orderings on X, I define the *majoritarian* and *Borda's* aggregation rules, denoted respectively by $\succsim^{maj}(\succsim_1, ..., \succsim_k)$ and $\succsim^{Bor}(\succsim_1, ..., \succsim_k)$, by:

[2] Since alternatives are mutually exclusive only one will be eventually chosen by some undefined tie-breaking mechanism.

$$x \succsim^{maj} (\succsim_1, ..., \succsim_k) y \iff \#\{i \leq k : x \succsim_i y\} \geq \#\{i \leq k : y \succsim_i x\}$$

$$x \succsim^{bor} (\succsim_1, ..., \succsim_k) y \iff \sum_{i=1}^{k} \#\{z \in X : x \succ_i z\} \geq \sum_{i=1}^{k} \#\{z \in X : y \succ_i z\}$$

The Borda's aggregation rule applied to a collection of orderings always generates an ordering. The majoritarian aggregation rule generates a binary relation that is reflexive and complete but not necessarily transitive. However, the majority relation always generates a transitive binary relation whenever it aggregates a collection of dichotomous orderings, i.e.:

Theorem 2. *(Inada 1964) Let $(\succsim_1, ..., \succsim_k)$ be a collection of k dichotomous orderings on X. Then the majoritarian aggregation $\succsim^{maj} (\succsim_1, ..., \succsim_k)$ is an ordering on X.*

2.3 Results

I first establish that any ordering is equivalent to the majoritarian aggregation of a collection of *distinct*[3] dichotomous orderings:

Theorem 3. *$\succsim$ is an ordering on X if and only if there exists a collection $(\succsim_1, ..., \succsim_k)$ of k distinct dichotomous orderings on X such that $\succsim^{maj} (\succsim_1, ..., \succsim_k) = \succsim$.*

Proof. Given Theorem 2, only the second implication needs to be established. Distinguish three cases: (i) $m(\succsim) > 2$, (ii) $m(\succsim) = 2$, and (iii) $m(\succsim) = 1$. For the first case, take a dichotomous ordering $\succsim_1$ such that the top indifference class is I^1 and the bottom indifference class is $\cup_{2 \leq j \leq m(\succsim)} I^j$. Then, take a second dichotomous ordering $\succsim_2$ such that the top indifference class is $I^1 \cup I^2$ and the bottom indifference class is $\cup_{3 \leq j \leq m(\succsim)} I^j$. Proceeding this way, we arrive at a collection $(\succsim_1, ..., \succsim_{m(\succsim)-1})$ dichotomous orderings such that all $x \in I^1$ are in the top indifference class of $m(\succsim) - 1$ dichotomous orderings,

[3] I require the collection of orderings to be distinct because the results are trivial otherwise since (i) any ordering is equivalent to any aggregation of the same single ordering and (ii) any ordering is equivalent to any aggregation of a collection of orderings identical to it.

all $x \in I^2$ are in the top indifference class of $m(\succsim) - 2$ dichotomous orderings, and so on. Then, straightforward verification shows that the binary relation induced by the majoritarian aggregation of $(\succsim_1, ..., \succsim_{m(\succsim)-1})$ is the ordering $\succsim$. For the second case, the previous construction generates a single dichotomous orderings $\succsim_1$. Then, add to this dichotomous ordering a universally equivalent (dichotomous) ordering and denote it $\succsim_2$. It follows that the binary relation induced by the majoritarian aggregation of $(\succsim_1, \succsim_2)$ is the ordering $\succsim$. For the third case, consider all pairs (x, y) of $\succsim$. Then, for each of these pairs take two dichotomous orderings $\succsim_1$ and $\succsim_2$ such that $x \succ_1 y \sim_1 a_1 \sim_1 a_2 \sim_1 ... \sim_1 a_{n-2}$ and $y \succ_2 x \sim_2 a_1 \sim_2 a_2 \sim_2 ... \sim_2 a_{n-2}$ where $a_1, ..., a_{n-2}$ are the remaining alternatives of X. Proceeding this way, we arrive at a collection $(\succsim_1, ..., \succsim_k)$ of $k = n(n-1)$ dichotomous orderings such that these express opposing preferences and cancel each other with respect to any given pair of alternatives. Then, the binary relation induced by the majoritarian aggregation of $(\succsim_1, ..., \succsim_k)$ is the ordering $\succsim$.

□

The original content of Theorem 3 can be viewed as a variation of McGarvey's (1953) theorem.[4] McGarvey (1953) shows that it is possible to view *any* complete binary relation $\succsim$ as the majoritarian aggregation of a collection of $n(n-1)$ *linear orderings*. Here, I show that by requiring the binary relation $\succsim$ to be transitive a similar result holds for a collection of *dichotomous orderings*.[5] Dichotomous orderings benefit from a very palatable interpretation, and is the only domain restriction *defined with respect to each individual binary relation* that is sufficient for the majoritarian aggregation rule to be always transitive when the number of binary relations is not necessarily odd.[6]

[4]See e.g. Hollard and Breton (1996) and Gibson and Powers (2012) for extensions of McGarvey (1953).

[5]Note that this result is not a corollary of McGarvey (1953). Although the McGarvey's (1953) theorem implies that any ordering can be viewed as a majoritarian aggregation of a collection of linear orderings, it does not entail that any ordering can be viewed as a majoritarian aggregation of a collection of dichotomous orderings.

[6]See Inada (1969) and Sen and Pattanaik (1969) for the remaining domain restrictions that are sufficient (and necessary) for the majoritarian aggregation rule to be always transitive when the number

It is worth noting that it is *not* possible to view any complete binary relation (contrary to any ordering) as the majoritarian aggregation of a collection of dichotomous preferences. For instance, suppose that $X = \{x, y, z\}$ and that $x \succ y$, $y \succ z$, and $z \succ x$. It is easy to check that there is no collection of dichotomous orderings such that a majoritarian aggregation induces this complete binary relation. The interested reader may also notice that the number of binary relations constructed in Theorem 3 is considerably lower than that in McGarvey (1953).[7] This points to the fact that the required number of *non-linear* orderings necessary for an ordering to be represented by the majority rule may be lower than the required number of *linear* orderings needed for an ordering (or a complete binary relation) to be represented by the same method. Now, I establish a similar result with respect to the Borda's aggregation rule:

Theorem 4. *$\succsim$ is an ordering on X if and only if there exists a collection $(\succsim_1, ..., \succsim_k)$ of k distinct linear orderings on X such that $\succsim^{bor}(\succsim_1, ..., \succsim_k) = \succsim$.*

Proof. To prove the non-trivial implication, distinguish two cases: (i) $m(\succsim) < n$ and (ii) $m(\succsim) = n$. For the first case (non-linear ordering), take two linear orderings $\succsim_1$ and $\succsim_2$ such that (1) for all $x \in I^j$ and all $y \in I^l$ with $1 \leq j < l \leq m(\succsim)$ one has $x \succ_1 y$ and $x \succ_2 y$, and (2) for all $x_p \in I^j$ with $p = 1, .., q$ one has $x_1 \succ_1 x_2 \succ_1 ... \succ_1 x_q$ and $x_q \succ_2 x_{q-1} \succ_2 ... \succ_2 x_1$. Since $m(\succsim) < n$, (2) guarantees that the two linear orderings are distinct. Then, straightforward verification shows that the binary relation induced by the Borda's aggregation of $(\succsim_1, \succsim_2)$ is the ordering $\succsim$. For the second case (linear ordering), consider all pairs (x, y) of $\succsim$. Then, for each of these pairs such that $x \succ y$ take two linear orderings $\succ_1$ and $\succ_2$ such that $x \succ_1 y \succ_1 a_1 \succ_1 a_2 \succ_1 ... \succ_1 a_{n-2}$ and $a_{n-2} \succ_2 a_{n-3} \succ_2 ... \succ_2 a_1 \succ_2 x \succ_2 y$; For each of these pairs such that $x \sim y$ take two linear orderings $\succ_1$ and $\succ_2$ such that $x \succ_1 y \succ_1 a_1 \succ_1 a_2 \succ_1 ... \succ_1 a_{n-2}$ and $a_{n-2} \succ_2 a_{n-3} \succ_2 ... \succ_2 a_1 \succ_2 y \succ_2 x$ where $a_1, ..., a_{n-2}$ are the remaining alternatives of

of binary relations is not necessarily odd. Contrary to dichotomous orderings, these restrictions are defined with respect to the admissible profile of binary relations.

[7] This is strictly true except when $m(\succsim) = 1$, since in that case the number of binary relations constructed is the same. See e.g. Stearns (1959), Deb (1976), and Brams and Fishburn (2002) for results and discussions concerning the minimal number of binary relations necessary to express any complete binary relation over a set made of n alternatives.

X. Proceeding this way, we arrive at a collection $(\succsim_1, ..., \succsim_k)$ of $k = n(n-1)$ linear orderings such that two of these express $\succsim$'s strict preference or indifference between any given pair of alternatives while all other linear orderings cancel with respect to this pair. Then, the binary relation induced by the Borda's aggregation of $(\succsim_1, ..., \succsim_k)$ is the ordering $\succsim$.

□

This result shows that an ordering is formally equivalent to the Borda's aggregation of a collection of distinct preferences. The interested reader may notice that whenever the binary relation is not a linear ordering, then one needs only two preferences to induce it by the Borda's rule. In a recent paper, Kelly and Qi (2016) show that for a fixed $k \geq 2$ any ordering is in the range of the Borda's rule except when k is odd and n is even. This subsumes Theorem 4 with the exception of the case in which $\succsim$ is a linear ordering (and since I look at a collection of distinct preferences).

These theorems show the *equivalence* between a single preference and two aggregations of a collection of preferences: it is possible to generate any single preference from these aggregations and these aggregations always generate a single preference. Then, it follows that the maximization of a single preference is not observationally distinguishable from these aggregations. This implication is captured in the following proposition that is a consequence of Theorems 3-4 and Arrow (1959):

Proposition 1. *The following statements are observationally equivalent:*

(i) A choice correspondence C is rationalized by the maximization of an ordering $\succsim$.

(ii) A choice correspondence C is rationalized by the majoritarian aggregation $\succsim^{maj}$ $(\succsim_1, ..., \succsim_k)$ of a collection $(\succsim_1, ..., \succsim_k)$ of k distinct dichotomous orderings.

(iii) A choice correspondence C is rationalized by the Borda's aggregation $\succsim^{Bor}$ $(\succsim_1, ..., \succsim_k)$ of a collection $(\succsim_1, ..., \succsim_k)$ of k distinct linear orderings.

(iv) A choice correspondence C satisfies WARP.

2.4 Discussion

Whenever a choice correspondence satisfies WARP (or any equivalent axiom), the results of this chapter show that the choice behavior can not only be rationalized by the direct maximization of a single preference but also by two aggregations of a larger collection of preferences. This means that these decision making models are observationally equivalent (Proposition 1).[8] Proposition 1 relates with the topics of (i) "indistinguishability" of two (or more) models and of (ii) model selection. In a recent paper, Manzini and Mariotti (2014) discuss the relation of these two issues in an interesting fashion. In what follows, I relate my results and arguments to their discussion.

According to Manzini and Mariotti (2014), two (or more) models are said to be indistinguishable if they are equivalent in terms of their empirical choice content, i.e., if they are characterized by the same observable conditions. There is by now some examples of this kind of indistinguishability in the literature. For instance, Moulin (1985) shows that any choice correspondence that is rationalized by a complete and quasi-transitive binary relation is also rationalized by a Pareto aggregation of a collection of linear orderings. More recently, Mandler, Manzini and Mariotti (2012) show that a decision making model based on the consecutive elimination of alternatives according to a sequence of desirable properties (a "checklist" model) is observationally equivalent in terms of choice to the maximization of a single preference.[9] In a more applied setting, Bernheim, Fradkin and Popov (2011) show that different behavioral explanations for the selection of a default option in pension plans are observationally equivalent in terms of choice.

[8] Note that the results of this chapter do not question the possibility to falsify the hypothesis that some observable behavior results from the maximization of a single preference; whenever a choice correspondence violates WARP it is possible to exclude that this behavior results from the maximization of a single preference (at least in principle: see e.g. Hausman 2000). In this respect, these results only show that it is also possible to exclude that the behavior that violates WARP results from two aggregations of a collection of preferences.

[9] For any subset A, an agent choosing according to a checklist model whittles down the set of eligible alternatives by first discarding those that do not possess the first *property* (that is defined as a set of alternatives), then those that do not possess the second property, and so on until one alternative remains which is the one that is chosen.

Manzini and Mariotti (2014, 351) argue that indistinguishability has significant implications for normative analysis when concerning individual decision making models; they argue that even if "there is no choice-based falsification of a particular model, we may still not be justified in using that model (rather than some other model also compatible with the choice data) as a basis for normative judgments". The authors give an example based on Mandler et al.'s (2012) result on the impossibility to distinguish the maximization of a single preference and a checklist decision making model. They argue that if choice behavior satisfies the revealed preference axioms, and those choices are the result of decision making by a checklist model (instead of the maximization of a single preference), then the standard revealed preference relation may not be a proper representation of the agent's welfare. For example, according to the checklist model an agent can reveal a preference for an alternative x over an alternative y exclusively on the basis that x possesses the first property while y does not (see Manzini and Mariotti 2014, 353). While it is possible that the agent does not care about the other properties, it is also possible that she cares about them but has not paid attention to them in order to limit the time and effort associated with choosing. Manzini and Mariotti (2014, 353) argue that if it is the case that y possesses *all* the other properties (besides the first) while x does not, that "it is dubious whether x can be declared welfare superior to y by a planner who has the time and resources to pay careful considerations to *all* aspects relevant for choice". They conclude that "[f]or a proper welfare inference, a considered judgement, involving an evaluation of the properties and their importance to the agent, is needed". I agree to a certain extent, and I will try to motivate why (and to what extent) in relation to the results of this chapter.

The results of this chapter lay open the question if indistinguishability may be an issue between individual and collective decisions (or between different collective models). By themselves, they seem to provide rather thin evidence on this direction. Though theoretically indistinguishable, the collective models, in particular due to their restriction in terms of the number of orderings needed to generate a single

preference, may not be plausible explanatory models for many contexts. As it will be argued next, questions regarding the plausibility of the different explanatory models need to be asked before treating indistinguishability as problematic. But other results also show that the behavior of organizations or teams that seem to decide on a collective basis may be observationally equivalent to "autocratic" decision making within an organization. For example, Moulin's (1985) result shows that it is not possible to distinguish if the choice behavior of an organization is the result of the decision of a single individual that is not able to compare indifferent alternatives *or* the result of a collective decision based on the Pareto aggregation of the preferences of its members. Similarly, Mandler et al.'s (2012) checklist model can be interpreted as a collective decision based on a *serial dictatorship*, in which for any subset A a first "dictator" selects a subset $B \in A$ of alternatives, from which a second "dictator" can choose a subset $C \in B$, and so on until one alternative is singled out to be chosen. Finally, indistinguishability between individual and collective decision making can also result from the fact that a "team" may choose consistently with the revealed preference axioms even if *their* preference is not the result of the aggregation of the individual preferences of the members of the team.[10] When taken together, these results suggest that it *may* be difficult to infer the underlying model of the choice behavior of an organization. This, as it was the case for individual decision making, *may* create difficulties to infer the individual welfare of the members of an organization based on revealed preference data.

But under which circumstances this may be the case? And even though the arguments related with welfare inference suggest that it may be important to know the underlying model that is behind a given choice behavior, when should we care? And if we have reasons to care, what to do when it is impossible to distinguish two (or more) decision making models based on revealed preference data? Which model to select?

[10] See e.g. Sugden 2000 for an essay on team preferences in this sense.

It seems that, as already hinted above, one should care only if at least two competing explanatory and indistinguishable models of a given choice behavior are *plausible* for the context of interest. According to Nooteboom (1986), a model is said to be "plausible" if it is composed of propositions/assumptions that are well connected and coherent with established knowledge. Let this knowledge concern a given context of interest. Then, plausibility can be checked at least in two ways. First, one can inquiry through introspection or casual observation about the connection/coherence between the model and the contexts where it is being applied/tested. For instance, there may be cases of clear autocratic decision making where it is not plausible to assume that a collective decision making model was behind such choices. Second, plausibility can be tested empirically through the collection of contextual information such as the timing of the decisions, frames, or the complexity of decisions. For instance, information on the complexity of a decision taken by a household may help us to include/exclude the participation of a child to that decision.

A second criterion to know when to care about indistinguishability seems to be if identifying the underlying model of some choice behavior is important. This may not always be the case. For example, the prediction of consumption behavior based on the traditional revealed preference axioms may be sufficiently accurate (for a given purpose) no matter what is the underlying model of the behavior. But in other cases it may be relevant to know the underlying model of some choice behavior. I have suggested that it may be important to distinguish between different models for the welfare inference of the members of an organization. Identifying the underlying model of observable behavior may be also important, for instance, when one wishes to understand and describe the behavior within an organization.

When two plausible and indistinguishable explanatory models are available to explain some observed behavior for which it is important to know its underlying decision making model, what can one do in order to identify the model that underlay choice behavior? One possibility, suggested by Manzini and Mariotti (2014), among many others, is to resort to (subjective) *non-choice data* such as verbal evidence,

survey data, psychological tests, and the like. According to Manzini and Mariotti (2014, 350-3) non-choice data should be subsidiary and referred to only in the case of a choice behavior that can be explained by several different explanatory models. We have seen that one can qualify this assertion for plausible explanatory models in contexts where it is important to know the underlying model of choice behavior.

Coming back to the welfare example given by Manzini and Mariotti (2014, 353), one would use non-choice data to distinguish between the maximization of a single preference and a checklist model only if both were plausible for the context in question and we were confronted with a situation for which it is important to identify the underlying model of choice behavior. For example, I conjecture that for a planner that do not wishes to "overrule" the agent's choice behavior it would not be important to identify which of the two models underlays the behavior. Both would give the same welfare relation. On the contrary, I conjecture that it would be important to distinguish them for a planner that do not wishes to overrule the agent's "self-authenticated" preferences. As argued in Chapter 1, one can use non-choice data to have information on the agent's *own* evaluation of her choices. In retrospect, the agent may state that the choice of x over y (as in the example above) was based on a checklist model and due to limited attention it was a mistake that does not reflect her preferences; alternatively, the agent may state that the choice reflected her "authentic" or reflexive preferences. This would, contrary to a choice-based criterion, possibly lead to different welfare relations depending upon the model that was adopted.

Though non-choice data may, in my view, take more than a subsidiary role in some economic applications (e.g. as in Benjamin, Heffetz, Kimball and Szembrot 2014), one should note that it does not solve the issue of indistinguishability when choice is the primary data. For one, subjective non-choice data may be unavailable. But even if available, non-choice data may fail to identify the underlying model behind some choice behavior. For example, some agents may imperfectly recall the process through which they arrived to a decision. Similarly, some agents may lie about their

motivations or decision process. If these issues will depend upon the context, subject-pool, the content of the questions or psychological tests, and the like, they should always be taken seriously when collecting this type of data.

Finally, if contextual information and non-choice data are inconclusive/unavailable, Manzini and Mariotti (2014) favor the *parsimony* of a model as an alternative criterion to rank competing explanations of choice. One can trace back such view at least to Friedman (1953). The author argues that in case of a tie on the criterion of predictive success, more parsimonious theories or theories that apply to a wider range of phenomena are to be preferred.[11] Manzini and Mariotti (2014, 353) argue that the parsimony criterion is severely under-used in economics. They also suggest, prudently, that "[s]ome measures [of parsimony] might be appropriate in some contexts and others in different contexts". Nevertheless, the criterion of parsimony seems at odds with the purpose of identifying the underlying model of choice behavior. Indeed, parsimony seems to be independent (and sometimes possibly opposed) to this end. An open question remains concerning what other criteria or methods to favor when the aim of identifying the underlying model of a general pattern of behavior is to be preserved.

2.5 Conclusion

The first message of this note can be summarized in one sentence: a single rationalization may hide multiple rationalizations. In effect, when one observes a choice correspondence that can be rationalized by a single preference, one cannot exclude that this rationalization results in fact from a (majoritarian or Borda) aggregation of a larger collection of preferences. The second message is that this will not always be problematic. Questions concerning the plausibility of the different explanatory models and if it is important to identify the underlying model of choice behavior need

[11] Remark that Friedman (1953) would consider the indistinguishability results of this chapter as favoring the maximization of a single preference as a good explanatory model, since they raise the scope of this theory.

to be asked. But given the theoretical prominence to favor choice data in economics (e.g. Gul and Pesendorfer 2008; Binmore 2009) and the increasingly application of revealed preference theory in fields such as household economics (e.g. Browning and Chiappori 1998; Cherchye, Rock and Vermeulen. 2010), these remarks highlight the relevance of pausing and ask some questions before committing to an interpretation of this type of data: won't it be *the tree that hides the forest*[12].

[12]I wish to thank Ilia Gouaref for reminding me of this proverb.

Bibliography

Afriat, S. (1967) The Construction of Utility Functions from Expenditure Data. International Economic Review 8(1): 67–77.

Archibald, G. C., H. A. Simon, and P. A. Samuelson (1963) Discussion. The American Economic Review: Papers and Proceedings 53(2): 227–36.

Arrow, K. J. (1959) Rational Choice Functions and Orderings. Economica 26: 121–27.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and N. Szembrot (2014) Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference. The American Economic Review 104(9): 2698–735.

Bernheim, B. D., A. Fradkin, and I. Popov (2011) The Welfare Economics of Default Options: A Theoretical and Empirical Analysis of 401 (k) Plans. , National Bureau of Economic Research Working Paper.

Binmore, K. (2009) Rational Decisions. Princeton University Press, Princeton, N.J.

Brams, S. and P. Fishburn (2002) Voting Procedures. In: K. Arrow A. K. Sen and K. Suzumura (eds) Handbook of Social Choice and Welfare. Elsevier Science, Amsterdam: 173–236.

Browning, M. and P.A. Chiappori (1998) Efficient Intra-household Allocations: A General Characterization and Empirical Tests. Econometrica 66(6): 1241–78.

Cherchye, L., B. De Rock, and F. Vermeulen. (2010) An Afriat Theorem for the Collective Model of Household Consumption. Journal of Economic Theory 145(3): 1142–63.

Deb, R. (1976) On Constructing Generalized Voting Paradoxes. The Review of Economic Studies 43(2): 347–51.

Friedman, M. (1953) The Methodology of Positive Economics. In: D. M. Hausman (ed) The Philosophy of Economics: An Anthology. Cambridge University Press, Cambridgethird edition (2008) edition.

Gibson, L. R. and R. C. Powers (2012) An Extension of McGarvey's Theorem from the Perspective of the Plurality Collective Choice Mechanism. Social Choice and Welfare 38(1): 101–08.

Gul, F. and W. Pesendorfer (2008) The Case for Mindless Economics. In: A. Caplin and A. Schotter (eds) The Foundations of Positive and Normative Economics. Oxford University Press, New York: 3–39.

Hausman, D. M. (2000) Revealed Preference, Belief, and Game Theory. Economics and Philosophy 16: 99–115.

Hollard, G. and M. Le Breton (1996) Logrolling and a McGarvey Theorem for Separable Tournaments. Social Choice and Welfare 13: 451–55.

Houthakker, H. (1950) Revealed Preference and the Utility Function. Economica 17: 159–74.

Inada, K. (1964) A Note on the Simple Majority Decision Rule. Econometrica 32(4): 525–31.

——— (1969) The Simple Majority Decision Rule. Econometrica 37(3): 490–506.

Kelly, J. S. and S. Qi (2016) A Conjecture on the Construction of Orderings by Borda's Rule. Social Choice and Welfare: 1–13.

Mandler, M., P. Manzini, and M. Mariotti (2012) A Million Answers to Twenty Questions: Choosing by Cheklist. Journal of Economic Theory 147: 71–92.

Manzini, P. and M. Mariotti (2014) Welfare Economics and Bounded Rationality: The Case for Model-based Approaches. Journal of Economic Methodology 21(4): 343–60.

McGarvey, D. C. (1953) A Theorem on the Construction of Voting Paradoxes. Econometrica 21(4): 608–10.

Moulin, H. (1985) Choice Functions Over a Finite Set: A Summary. Social Choice and Welfare 2: 147–60.

Nooteboom, B. (1986) Plausibility in Economics. Economics and Philosophy 2: 197–224.

Richter, M. K. (1966) Revealed Preference Theory. Econometrica 34: 635–45.

Samuelson, P. A. (1938) A Note on the Pure Theory of Consumer's Behaviour. Economica 5(17): 61–71.

Sen, A. K. (1971) Choice Functions and Revealed Preferences. Review of Economic Studies 38: 307–17.

Sen, A. K. and P. K. Pattanaik (1969) Necessary and Sufficient Conditions for Rational Choice under Majority Decision. Journal of Economic Theory 1: 178–202.

Stearns, R. (1959) The Voting Problem. The American Mathematical Monthly 66(9): 761–63.

Sugden, R. (2000) Team Preferences. Economics and Philosophy 16: 175–204.

Chapter 3

Choice with Time*

We propose a framework for the analysis of choice behavior when the later explicitly depends upon time. We relate this framework to the traditional setting from which time is absent. We illustrate the usefulness of the introduction of time by proposing three possible models of choice behavior in such a framework: (i) changing preferences, (ii) preference formation by trial and error, and (iii) choice with endogenous *status-quo* bias. We provide a full characterization of each of these three choice models by means of revealed preference-like axioms that could not be formulated in a timeless setting.

Keywords: Choice, behavior; Time; Revealed preferences; Changing preferences; Learning by trial-and-error; Inertia bias.

3.1 Introduction

An important accomplishment of modern economic theory is the precise identification of its behavioral implications. A rich - and now classical - tradition of research, initiated by Samuelson (1938), and pursued by Houthakker (1950), Chernoff (1954), Arrow (1959), Richter (1966), Sen (1971), among many others, has formulated these implications in terms of a *choice function*, sometimes generalized to a *choice*

*This chapter is adapted with some liberty from a joint work with Nicolas Gravel.

correspondence. While a choice function assigns to every set of alternatives - or menu - in some universe a unique element of it, interpreted as the chosen alternative in the menu, a choice correspondence assigns to every menu in the universe a *subset* of this menu, interpreted as containing all alternatives that *could have* been chosen by the agent. A choice correspondence is not directly observable because we do not in practice observe simultaneous multiple choices over complete and mutually exclusive descriptions of the world. For this reason, we focus mainly on choice functions in what follows.

The behavioral implications of a significant variety of theories have been examined through the formalism of choice functions. The most well-known of them posits that the choice results from the maximization of a single preference defined on the set of all conceivable alternatives. The behavioral implications of this theory on abstract choice functions are the Chernoff (1954) condition (called property α by Sen 1971), Arrow's (1959) condition, Houthakker's (1950) axiom of revealed preference or the Richter's (1966) congruence axioms. This one-rationale explanation of choice has also been applied to the specific context of classical consumer theory where the alternatives are consumption bundles and where the menus are budget sets. In this setting, where additional properties of preferences such as local non-satiation can be defined, the most well-known empirical implication of this one-rationale choice theory is Afriat (1967) Generalized Axiom of Revealed Preferences (GARP), very clearly analyzed in Varian (1982) and discussed in Varian (2006).

The findings of psychology and behavioral economics suggest, however, that the implications of the maximization of a single preference are often rejected by actual choice behavior (see e.g. Fudenberg 2006, Fehr and Hoff 2011, and Hoff and Stiglitz 2016 for reviews). This has led several authors to propose alternative theories of choice and to look for the implications of these on the choice function or correspondence. For example, Masatlioglu et al. (2012) have identified the properties of a choice function that selects the preferred alternative from a *consideration set* in each menu, rather than from the whole menu itself. This consideration set is interpreted

as reflecting what the decision maker pays attention to in her choice process. This consideration set may not coincide with the whole menu of feasible alternatives if, for example, the decision maker is "inattentive" to some of the alternative that are available. Barberà and Neme (2016) have also used a choice function to characterize a model in which the decision maker chooses one of the r-best alternatives according to a preference, rather than the 1st-best as assumed in conventional theory. Sprumont (2000) has used a choice correspondence to identify the implications that a collection of individual agents choose an alternative that could be a Nash equilibrium of a game for some preferences. Others, such as Manzini and Mariotti (2007, 2012) and Apesteguia and Ballester (2013), have identified the observable properties of a choice function that are necessary and sufficient for its rationalization by a sequential lexicographic application of a collection of preferences that ends up selecting a unique alternative from each menu. These examples illustrate how a choice function can be used to test competing explanations for the observed choices of an economic agent that cannot be explained by the maximization of a single preference relation.

Flexible and amenable to formulations of testable implications of many (behavioral) choice models as it is, a choice function (or correspondence) may still be considered unduly abstract for many applications. One of the important and easily observable feature of the reality that it neglects is the *time* at which the menu is made available to the decision maker. Indeed, as used in the literature just described, a choice function describes a timeless process that only provides the chosen alternatives in every admissible menu. It does not record (nor use information on) the *periods* at which the menus are available. Yet, in most data on choice observations that we could think of, the menus of choice will present to the decision maker one after the other and this information is known. For instance, economic experiments often record (or can record) the time sequence of choices. More generally, dynamic choice theory provides several examples where the time sequence of choices plays a key role. Change of habits, learning, and similar phenomena in which the preferences

appear to be endogenous to the *experience* of the decision maker seem to require an explicit integration of time in the description of the choice process.

In this chapter, we therefore extend the traditional setting and propose to analyze the choice behavior as an explicit function of *both* the time at which the choice takes place *and* the menu available at that time. We then show that this information on the time period at which the choice takes place enables one to identify the behavioral implications of theories of choice that could not be analyzed without an explicit integration of time. Three examples of such theories are examined and characterized in this chapter. Each of the examples relates with one important literature in behavioral economics: preference reversals, learning, and cognitive reference-dependent bias.

The *first* one concerns the possibility, for the decision maker, to experience a *change in preferences* at some period. In such a model, the decision maker chooses in a way that maximizes a given preference up to some time period and, after this period, switches to another preference and makes its subsequent choices based on this preference. We provide a simple "revealed preference test" for this particular theory of choice, that relates to the literature on changing tastes (see e.g. Gul and Pesendorfer 2005). While we characterize a choice model in which a decision maker changes preferences only once, the generalization to any finite number of changes in preferences that is smaller than the total number of time periods would be straightforward.[1]

The *second* theory of choice that we characterize with a simple revealed preference axiom is that of *learning by trial and error*. In such a theory, the decision maker "tries out" the alternatives before forming her preference over them. Hence, when facing a menu at a given period, the decision maker either tries out one alternative *or* chooses the "best" option according to a single preference relation among the alternatives she has previously chosen at least once. This model provides a rationalization of "inconsistent" behavior at the beginning of some sequence of choices. It

[1]Trivially, any choice behavior that depends upon time can be seen as resulting from a decision maker who changes preferences at every period. See Kalai et al. (2002) for a similar observation on the standard timeless setting.

describes a plausible process of "trial and error" for discovering what a person "really prefers" that has been widely documented in the literature. It relates to Cooke (2016) and Piermont et al. (2016) who characterize similar learning models over uncertain prospects and to Young (2009) who characterizes a similar learning model in a game theoretical environment.

The *third* theory of choice characterized herein is what could be called *choice with inertia bias*. In such a theory, the decision maker takes her last choice as a *default* for her current choice. For each choice situation, the decision maker chooses the best option according to a single preference *or* the alternative she has chosen in the previous period. We interpret this resort to the choice made in the immediately previous period as an "imprisonment in the habit". There has been many contributions in the literature that have examined behavior exhibiting *status-quo* bias (for example Tversky and Kahneman 1991). In a related vein, several authors have provided axiomatic characterizations of choice models with a default-option in which this option is exogenously given (e.g. Bossert and Sprumont 2003; Masatlioglu and Ok 2005). In this theory of choice with inertia bias, the default-option is rather endogenous and evolves over time.

These examples are, of course, extremely specific representations of a much wider and richer class of choice models involving non-standard considerations such as preference reversals, learning, and cognitive reference-dependent bias. Yet, we believe that these examples serve rather well their purpose of showing how the introduction of time in the formal description of choice behavior is necessary for identifying the empirical implications of several theories of choice, and how it eases the identification of these implications. Though only the first model is consistent with a model of decision making based on multiple preferences, these examples also illustrate the potential of this framework to study other choice models motivated by endogenous and multiple preferences.

The framework introduced in this chapter bears some similarity with that introduced by Bernheim and Rangel (2007; 2009) and Salant and Rubinstein (2008).

These authors have analyzed some normative (Bernheim and Rangel 2007; 2009) and positive (Salant and Rubinstein 2008) implications of choice processes in which every menu of alternatives is supplemented with an *ancillary condition* that represents either a "frame" or some other "consequentially-irrelevant" feature of the choice environment. One could, of course, view the time at which the choice is made as a frame or an ancillary condition. Yet time is a somewhat specific feature of the choice environment. One of its specificities is that it leads to an ordering of the menus offered to the decision maker as *per* the time at which they are available. The properties of this ordering (e.g. the fact that one alternative chosen "in the past" is not chosen "in the present") play an important role in the characterization of the choice models that we provide. By contrast, the abstract ancillary conditions and frames examined by Bernheim and Rangel (2007; 2009) and Salant and Rubinstein (2008) do not impose a structure on the set of available menus that is as precise as that of an ordering. Another difference between our approach and those of Bernheim and Rangel (2007; 2009) and Salant and Rubinstein (2008) is that we do not assume the possibility of observing choice behavior that would take place in any conceivable combination of time period and menu at that time period. We only consider, somewhat realistically, that we observe a particular chronology of choices, and we identify the necessary and sufficient conditions that the choice behavior observed in that particular chronology must satisfy in order to result from each of the three theories of choice mentioned above. A third difference between the approach of Bernheim and Rangel (2007; 2009) and Salant and Rubinstein (2008) and ours concerns the interpretation given by the former to the frame and the ancillary condition. Bernheim and Rangel (2007; 2009) define an ancillary condition to be "a feature of the choice environment that may affect behavior, but [that] is not taken as relevant to a social planner's evaluation" (Bernheim and Rangel 2009, 55). As illustrated in Chapter 1, considering time as a feature of the choice environment that is irrelevant to a social planner's evaluation may create some difficulties in properly ranking different states of affairs. For example, our intuition suggests that preferences (choices) that

were revealed (made) very long time ago may have less bearing on our appraisal of the current well-being of the decision maker than those revealed (made) in more recent periods. Still, in what respects this chapter, nothing in our analysis depends on this intuition. Similarly, we also have difficulty in viewing the time at which a choice is made as a frame in Salant and Rubinstein's (2008, 1287) sense, i.e., as an information that is "irrelevant in the rational assessment of the alternatives but nonetheless affects behavior". We suspect that Salant and Rubinstein (2008), who contrary to Bernheim and Rangel (2007; 2009) do not give time as an example of a frame, would agree with us.

In a recent paper, Cerigioni (2016) proposes a choice theoretic framework that explicitly introduces time, but in which the menu available for choice at every time period is supplemented by an abstract vector of (non-time) ancillary conditions that themselves depend upon the time period. He characterizes in this framework a "dual-self" theory of choice. As compared to his, our analysis is therefore closer to the classical choice theory since, except for time, we do not consider any other argument of the choice function than the menu of alternatives to which it applies.

The remainder of the chapter is organized as follows. The next section is devoted to the formal introduction of our framework of choice with time, and we discuss its connection with the classical timeless choice theoretic setting. In Section 3.3 we characterize the traditional choice model of the maximization of a time-invariant preference, showing that the introduction of time is irrelevant for the rationalization of the standard theory of rational choice. In Section 3.4 we define the three time-dependent theories of choice discussed above and identify their implications on choice behavior. We conclude with a brief discussion in Section 3.5.

3.2 Framework

3.2.1 Choice Domain

Let X be a universe of alternatives of interest for the decision maker, let $\mathcal{P}(X)$ be the set of all non-empty subsets of X, and $\mathcal{F} \subseteq \mathcal{P}(X)$ be a collection of subsets of X, each of which being interpreted as a *choice problem* (using Arrow's 1959 terminology) or a *menu*. A choice function on $\mathcal{F}$ is a mapping $C : \mathcal{F} \longrightarrow X$ that satisfies $C(A) \in A$ for every A in $\mathcal{F}$. The choice-theoretic literature that has emerged in the last sixty years or so has made various assumptions on the domain $\mathcal{F}$ that depend, sometimes, upon the nature of the alternatives in X that are considered. For example, the classical theory introduced by Arrow (1959) has taken X to be an abstract finite set, and $\mathcal{F}$ to coincide with $\mathcal{P}(X)$. This is clearly very demanding from an observational viewpoint, since it is difficult in practice to observe all choices that an agent could make when facing any conceivable menu.

Quite a few years later, Sen (1971) has shown that several results on the rationalization of choice models hold on more restricted domains provided that those domains include all possible pairs and triples of X. While less demanding, this requirement is still quite demanding, as we often do not observe (or do not want to "have" to observe) all possible pairs and triples of X, even when the latter is finite. In another attempt to relax the observational demands of the classical theory, Richter (1966; 1971), Hansson (1968), and Suzumura (1976; 1977; 1983) developed the theory of revealed preference for choice functions or correspondences with **general domains** that do not impose any restriction on the class of menus that may be available. In particular, the authors provided several characterizations of choice functions and/or correspondences defined on any non-empty family of non-empty subsets of X.[2]

Note that the observational implications of a choice model that are found in this domain apply to Arrow's domain that include all possible choice problems. However, the same cannot be said in the opposite direction. For example, some of the necessary

[2]See also Bossert et al. (2005; 2006).

and sufficient conditions for some behavior to be rationalized by the maximization of a single preference in Arrow's domain are not necessary and sufficient in a general domain (Suzumura 1976, 155-6). Since - as in the cases of consumer behavior and choice-based experiments - we often do not observe all conceivable choice problems (neither all pairs and triples of X), the axioms found in Arrow's domain may not be necessary and sufficient for empirical applications. On the contrary, the "status of the general domain seems to be impeccable, as the theory developed on this domain is relevant in whatever choice situations we may care to specify" (Bossert et al. 2005, 186).

3.2.2 Choice behavior and time

In the following, we supplement this general domain with a discrete time horizon $\mathcal{T} = \{1, ..., T\}$. This enables one to define a **chronology of choices** as a function $A : \mathcal{T} \longrightarrow \mathcal{F}$ that assigns to every choice period $t \in \mathcal{T}$ a unique non-empty set $A(t) \in \mathcal{F}$, interpreted as the menu available at *period t*. A **chronological choice function** C for the chronology A is simply a mapping that assigns to every pair $(t, A(t))$ of that chronology a unique element $C(t, A(t)) \in A(t)$.

From a formal point of view, a chronological choice function has two arguments: the choice period at which the choice is made, and the menu available at that period. Because of this, it is possible to have $C(t, A(t)) \neq C(t', A(t'))$ even if $A(t) = A(t')$. That is, a decision maker who faces the same menu at two different time periods may make different choices in this menu. Such a possibility is of course ruled out in the classical timeless choice theoretic framework. On the other hand, since a chronology of choice is taken to be a function from $\mathcal{T}$ to $\mathcal{F}$, one cannot have two different menus available at the same period.

This formulation seems adapted for several purposes in economics. First, a chronological choice function can be used for the *ex-post* rationalization of some observed behavior. Second, it can also be used for the normative/welfare appraisal of observed

outcomes. Third, a chronological choice function seems also to be adapted for the prediction of some future behavior. By observing if a chronological choice function satisfies given behavioral consistency properties, one can impose consistency requirements upon choices that have not yet been observed in the time horizon. Finally, our framework seems adapted to extend the observed choices on some $\mathcal{F} \subset \mathcal{P}(X)$ on to the entire set $\mathcal{P}(X)$. In standard positive economics, this is done by estimating (when possible) a time-invariant preference from observed choices in some $\mathcal{F} \subset \mathcal{P}(X)$, and using this preference to infer choices in $\mathcal{P}(X) \setminus \mathcal{F}$. The same can be done in our setting, either for a time-invariant preference or other decision making models.

Just like in the standard timeless framework, the behavioral implications of the theories that we are looking after will take the form of axioms that are formulated in terms of "revealed preference" relations. For now, two types of such relations shall be considered. The first one is the *direct revealed preference* relation at time t that is defined as follows.

Definition 1. *For any period $t \in \mathcal{T}$ and alternatives x and $y \in X$, we say that x **is directly revealed preferred to** y **at period** t, denoted $x \succsim_C^t y$, if and only if $x = C(t, A(t))$ and $y \in A(t)$.*

In words, the chronological choice function directly reveals a preference for x over y at period t (with x and y distinct) whenever it shows the choice of x at period t in a choice problem where y was available. This direct revealed preference at period t is analogous to the notion formulated by Arrow (1959) in a timeless setting. In the spirit now of Houthakker (1950), one can define the notion of *indirect revealed preference* relation over a time period going from r to s as follows:

Definition 2. *For any periods r and $s \in \mathcal{T}$ such that $r \leq s$ and any alternatives x and $y \in X$, we say that x **is indirectly revealed preferred to** y **between periods** r **and** s, denoted $x \succsim_C^{rs} y$, if and only if there is a sequence $\{t_j\}_{j=1}^k$ of k time periods in the set $\{r, r+1, ..., s-1, s\}$, not necessarily ordered by time, for which one has:*
(i) $x = C(t_1, A(t_1))$,

(ii) $C(t_j, A(t_j)) \succsim_C^{t_j} C(t_{j+1}, A(t_{j+1}))$ *for all* $j = 1, ..., k-1$, *and*
(iii) $y \in A(t_k)$.

We observe that, by the very definition of a chronological choice function, the binary relation $\succsim_C^t$ is antisymmetric for every period t. However, for an arbitrary pair of periods r and s satisfying $r \leq s$, the binary relation $\succsim_C^{rs}$ need not be antisymmetric. The fact of having $x \succsim_C^{rs} y$ for two distinct alternatives x and y does not preclude the possibility of having $y \succsim_C^{rs} x$. We emphasize that the sequence of sets involved in the definition of the indirect revealed preference between periods r and s need not be ordered by time. To give just an example, suppose that $r = 1$, $s = 3$, $X = \{a, b, c, d, e\}$, that the chronology of choices offered to the decision maker between period 1 and period 3 is $A(1) = \{a, b, c\}$, $A(2) = \{d, a\}$ and $A(3) = \{b, e\}$, and that the chronological choice function for the periods 1, 2 and 3 is:

$$\begin{aligned} C(1, A(1)) &= a, \\ C(2, A(2)) &= d, \text{ and} \\ C(3, A(3)) &= b. \end{aligned}$$

If follows from Definition 2 that alternative d is indirectly revealed preferred between periods 1 and 3 to alternative e. In effect, d has been directly revealed preferred to a in period 2 which has been itself directly revealed preferred to b in period 1 which has been directly revealed preferred to e at period 3. The sequence of direct revealed preference statements that connect d to e between periods 1 and 3 is not indexed by time.

3.3 The Single Preference Choice Model and Time

Standard choice theory is grounded on the assumption that preferences, and the choices that they induce, are invariant with respect to time. If empirical evidence, casual observation, and introspection suggest that this assumption is not always real-

istic, it does represent a sensible benchmark for many applications. We therefore find it useful to start our analysis by characterizing a chronological choice function that results from the maximization of a (linear) ordering. The axiom that characterizes this behavior is the following.

Axiom 1. *For any periods r, s and t such that $r \leq s < t$ and some distinct x and $y \in X$, one cannot have $x \succsim_C^{rs} y$ and $y \succ_C^{t} x$.*

We note that this axiom is somewhat simpler to test than GARP. In effect, Axiom 1 requires a consistency between indirect revealed preferences occurring between any two periods r and s, and direct revealed preferences expressed at subsequent time period t. By contrast, the standard timeless GARP test would have ruled inconsistencies also between indirect revealed preferences occurring between any two periods r and s and any direct revealed preference whatsoever, including those observed before r. The later test is then computationally slightly more demanding.

We now define what it means for a chronological choice behavior to result from the maximization of a time-invariant preference.

Definition 3. *A chronological choice function C results from the maximization of a time-invariant preference if and only if there exists a linear ordering $\succsim$ on X such that, for every $t \in \mathcal{T}$, one has $a = C(t, A(t))$ if and only if $a \succsim a'$ for all $a' \in A(t)$.*

We now establish that Axiom 1 is necessary and sufficient for a chronological choice function to result from the maximization of a time-invariant preference.

Theorem 1. *A chronological choice function C satisfies Axiom 1 if and only if it results from the maximization of a time-invariant preference.*

Proof. We first show that a chronological choice function C for which there exists a linear ordering $\succsim$ on X such that, for every $t \in \mathcal{T}$, one has $a = C(t, A(t))$ if and only if $a \succsim a'$ for all $a' \in A(t)$ satisfies Axiom 1. For this sake, assume the existence of a

linear ordering $\succsim$ on X such that, for every $t \in \mathcal{T}$, one has $a = C(t, A(t))$ if and only if $a \succsim a'$ for all $a' \in A(t)$ and consider any periods r, s and t such that $r \leq s < t$ and some distinct x and $y \in X$ for which we have $x \succsim_C^{rs} y$. By Definition 2 , there is a sequence $\{t_j\}_{j=1}^k$ of k time periods in the set $\{r, r+1, ..., s-1, s\}$ for which one has $x = C(t_1, A(t_1))$, $C(t_j, A(t_j)) \succsim_C^{t_j} C(t_{j+1}, A(t_{j+1}))$ for all $j = 1, ..., k-1$, and $y \in A(t_k)$. Since the chronological choice function is rationalized by the linear ordering $\succsim$, one has $C(t_j, A(t_j)) \succsim C(t_{j+1}, A(t_{j+1}))$ for all $j = 1, ..., k-1$ and, therefore, $x \succsim y$ by the transitivity of $\succsim$. Assume contrary to Axiom 1 that $y \succ_C^t x$. By Definition 1, this means that $y = C(t, A(t))$ and $x \in A(t)$. Since $\succsim$ rationalizes the chronological choice function C, this means that $y \succ x$, a contradiction.

To prove the other implication, consider a chronological choice function C that satisfies Axiom 1. Define the binary relation $\succsim_C$ on X by:

$$x \succsim_C y \iff \exists t \in \mathcal{T} \text{ s.t. } x = C(t, A(t)) \text{ and } y \in A(t).$$

Define also the binary relation $\widehat{\succsim}_C$ by:

$$\begin{aligned} x \,\widehat{\succsim}_C\, y \quad &\iff \quad \exists \{t_j\}_{j=0}^k \text{ with } t_j \in \mathcal{T} \text{ for } j = 0, ..., k \text{ and } l \geq 0) \text{ such that} \\ \text{(i) } x \quad &= \quad C(t_0, A(t_0)) \\ \text{(ii) } C(t_{j+1}, A(t_{j+1})) \quad &\in \quad A(t_j) \text{ for } j = 0, ..., l-1 \text{ (if any)} \\ \text{(iii) } y \quad &\in \quad A(t_l) \end{aligned}$$

It is immediate to see that $\widehat{\succsim}_C$ is the transitive closure of $\succsim_C$. This means that $\widehat{\succsim}_C$ is transitive by definition. We now show that $\widehat{\succsim}_C$ is antisymmetric if C satisfies Axiom 1. By contradiction, suppose $\widehat{\succsim}_C$ is such that there are two distinct alternatives x and

$y \in X$ for which both $x \widehat{\succsim}_C y$ and $y \widehat{\succsim}_C x$ holds. This means that:

$$\begin{aligned} \exists\{t_j\}_{j=0}^{k} \text{ with } t_j &\in \mathcal{T} \text{ for } j = 0, ..., k \text{ (with } k \geq 0) \text{ such that} \\ \text{(i) } x &= C(t_0, A(t_0)) \\ \text{(ii) } C(t_{j+1}, A(t_{j+1})) &\in A(t_j) \text{ for } j = 0, ..., l-1 \text{ (if any)} \\ \text{(iii) } y &\in A(t_l) \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} \exists\{t'_j\}_{j=0}^{l'} \text{ with } t'_j &\in \mathcal{T} \text{ for } j = 0, ..., l' \text{ (with } l' \geq 0) \text{ such that} \\ \text{(i) } y &= C(t'_0, A(t'_0)) \\ \text{(ii) } C(t'_{j+1}, A(t'_{j+1})) &\in A(t'_j) \text{ for } j = 0, ..., l'-1 \text{ (if any)} \\ \text{(iii) } x &\in A(t_{l'}) \end{aligned} \tag{3.2}$$

Consider the sets of time periods $T = \bigcup_{j=0}^{l} \{t_j\}$ and $T' = \bigcup_{j=0}^{l'} \{t'_j\}$ involved in expressions (3.1) and (3.2) respectively. As these two expressions define a cycle of revealed preference relations connecting alternative x to itself, this cycle can be started at any $C(t, A(t))$ (for some $t \in T \cup T'$) that we wish. In particular, t can be the maximal (with respect to the natural ordering of time) such period in $T \cup T'$. We then have $C(t, A(t)) \succsim_C^t C(s, A(s))$ for some $s < t$. By definition of the cycle induced by expressions (3.1) and (3.2), there is also a period $r < t$ such that $C(t, A(t)) \in A(r)$. Let (r, s) denote the set of all choice problems between r and s, and let $s' = \max(r, s)$. Using the definition of the cycle and Definition 2, it follows that $C(s, A(s)) \succsim_C^{1s'} C(t, A(t))$ and $C(t, A(t)) \succsim_C^t C(s, A(s))$, a contradiction of Axiom 1. Hence $\widehat{\succsim}_C$ is an antisymmetric and transitive binary relation. By Spilrajn extension theorem, one can therefore extend $\widehat{\succsim}_C$ into a complete linear ordering $\succsim$. Let us now show that for every $t \in \mathcal{T}$, one has $x = C(t, A(t)) \Longleftrightarrow x \succsim a$ for all $a \in A(t)$. Consider therefore any $t \in \mathcal{T}$. Assume first $x = C(t, A(t))$. Then, by definition of $\succsim_C$, one has $x \succsim_C a$ for every $a \in A(t)$ so that the implication $x \succsim a$ for every $a \in A(t)$ follows from the fact that $\succsim$ extends $\widehat{\succsim}_C$ which extends itself

$\succsim_C$. Assume now that $x \succsim a$ for every $a \in A(t)$ for some $x \in A(t)$. Suppose by contradiction that $x \neq C(t, A(t))$. Then, there exists some alternative y distinct from x such that $y = C(t, A(t))$. By definition of $\succsim_C$, one has $y \succsim_C x$ and, therefore, $y \,\widehat{\succsim}_C\, x$ and $y \succsim x$. But, since $x \succsim a$ for every $a \in A(t)$, this is incompatible with $\succsim$ being antisymmetric. □

Theorem 1 shows that Axiom 1 is necessary and sufficient for "rational behavior". At the same time, the proof of Theorem 1 clearly suggests that indexing the choice behavior by time is irrelevant for the possibility of rationalizing that behavior as resulting from the maximization of a linear ordering.

Even though the result of Theorem 1 is simple, we notice that, to the best of our knowledge, it has never been established before for a choice function defined on an arbitrary domain. On an arbitrary domain, Suzumura (1976) shows that SARP (a slight strengthening of GARP) is a necessary and sufficient condition for a *choice correspondence* to be rationalized by a complete and acyclical binary relation.

While the introduction of time does not play any significant role in characterizing one-preference rational behavior, we show in the next section that there are alternative theories of choice whose empirical implications cannot be characterized without an explicit inclusion of time.

3.4 Examples of Time-dependent Choice Models

There is a growing support to the view that the economic agent's preferences are best represented as *time-dependent* and that we often observe choice behavior for the same set of alternatives to differ across time. Preference (and choice) reversals, learning, and several types of cognitive bias are among the phenomena most studied in behavioral economics (see e.g. Fehr and Hoff 2011; Hoff and Stiglitz 2016).

In what follows, we provide characterizations of three choice models that relate with each of these phenomena: a model of changing preferences (Section 3.4.1), a model of preference formation through trial and error (Section 3.4.2), and a model

of endogenous *status-quo* bias due to inertia in preferences (Section 3.4.3). Though simple and not imposing strong structure upon choice behavior, they intend to illustrate the potential of this framework to study choice models motivated by multiple and *endogenous preferences* (i.e., models in which choice is endogenous to the experience of the decision maker).

3.4.1 Changing Preferences

The first of these models considers the possibility for the decision maker to behave as if her preferences or tastes were *unpredictably* changing over time. We analyze the case in which the decision maker preferences may change unpredictably *at most* once. From the observer point of view, this corresponds to the case where there is a single period, unknown *a priori* by the observer, in which the decision maker "switches" from one preference to another. For instance, someone that likes meat could become vegetarian at a given point in time.

To see the reach of this changing preferences model, and the relevance of including time in the analysis of its behavioral implications, consider the following two examples:

Example 1. *Let* $\mathcal{T} = \{1, 2, 3\}$ *and consider the following chronology of (gastronomic) menus:*

$A(1) = \{chicken, dahl\}$, $A(2) = \{chicken, dahl, tuna\}$, $A(3) = \{chicken, dahl, beef\}$.

The chronological choice function C *defined by* $C(1, A(1)) = chicken$, $C(2, A(2)) = dahl$, *and* $C(3, A(3)) = chicken$ *is not consistent with one change in preferences. Indeed, both the choices at the first and at the last period reveal a (carnivorous) preference for chicken over dahl, while the choice made at the second period reveals a preference for dahl over chicken. In order to generate such a pattern of choice, the decision maker must have changed preferences at every period.*

Example 2. *Let again consider* $\mathcal{T} = \{1, 2, 3\}$ *and the same chronology:*

$A(1) = \{chicken, dahl\}$, $A(2) = \{chicken, dahl, tuna\}$, $A(3) = \{chicken, dahl, beef\}$.

The chronological choice function C defined by $C(1, A(1)) = C(2, A(2)) =$ chicken and $C(3, A(3)) =$ dahl is consistent with one change in preferences. The decision maker switches once for a vegetarian preference between the second and the third periods.

Notice that it would not be possible to distinguish between the two choice behaviors without the introduction of time. Indeed, without time, both examples entail a single violation of GARP in the traditional sense.

We now provide an axiom on a chronological choice function that characterizes a decision maker who chooses in every period according to some preference relation, and who experienced at most one preference change in time.

Axiom 2. *If there are periods r, s and t such that $r \leq s < t$ and $x \succ_C^{rs} y$ and $y \succ_C^{t} x$ for some distinct x and $y \in X$, then, for every distinct w and $z \in X$, one cannot have $w \succsim_C^{uv} z$ and $z \succ_C^{\tau} w$ for periods u, v and τ such that $t \leq u \leq v < \tau$.*

This axiom says that if one observes a violation of Axiom 1 between period 1 and a given period t, then it is not possible to observe a second violation of Axiom 1 between t and T. This axiom is therefore almost as easy to test as Axiom 1. We now define what is meant by a chronological choice behavior to result from one change in preferences.

Definition 4. *A chronological choice function C results from one change in preferences if there exists two (possibly identical) linear orderings $\succsim_1$and $\succsim_2$ on X and one period $t \in \mathcal{T}$ such that $a_j = C(j, A(j))$ if and only if $a_j \succsim_1 a'_j$ for all $a'_j \in A(j)$ and $j \in \mathcal{T}$ such that $j < t$ and $a_v = C(v, A(v))$ if and only if $a_v \succsim_2 a'_v$ for all $a'_v \in A(v)$ and for all $v \in \mathcal{T}$ such that $v \geq t$.*

The characterization of this choice model is provided in the following theorem.

Theorem 2. *A chronological choice function C satisfies Axiom 2 if and only if it results from one change in preferences.*

Proof. For the necessity of the condition, assume that C is a chronological choice function that results from one change in preferences as per Definition 4. This means that there exists two (possibly identical) linear orderings $\succsim_1$and $\succsim_2$ on X and one period $t \in \mathcal{T}$ such that $a_j = C(j, A(j))$ if and only if $a_j \succsim_1 a'_j$ for all $a'_j \in A(j)$ and $j \in \mathcal{T}$ such that $j < t$ and $a_v = C(v, A(v))$ if and only if $a_v \succsim_2 a'_v$ for all $a'_v \in A(v)$ and for all $v \in \mathcal{T}$ such that $v \geq t$. Assume by contradiction that this chronological choice function violates Axiom 2. That is, assume that there are periods r, s and t' satisfying $r \leq s < t'$ for which one has $x \succ_C^{rs} y$ and $y \succ_C^{t'} x$ for some distinct x and $y \in X$, and that there are also some distinct w and $z \in X$ for which one observes $w \succsim_C^{uv} z$ and $z \succ_C^{\tau} w$ for some periods u, v and τ such that $t' \leq u \leq v < \tau$. We first show that having both $x \succ_C^{rs} y$ and $y \succ_C^{t'} x$ implies that $r < t \leq t'$. By contradiction, suppose first that $t \leq r < t'$. Since one has $a_v = C(v, A(v))$ if and only if $a_v \succsim_2 a'_v$ for all $a'_v \in A(v)$ and all v such that $v \geq t$, the fact of observing both $x \succ_C^{rs} y$ and $y \succ_C^{t'} x$ would imply, given the definition of $\succ_C^{rs}$and $\succ_C^{t'}$ and the transitivity of $\succsim_2$, that both $x \succ_2 y$ and $y \succ_2 x$ holds, which is a contradiction. Similarly, if $r < t' < t$, and given the fact that $a_j = C(j, A(j))$ if and only if $a_j \succsim_1 a'_j$ for all $a'_j \in A(j)$ and all j such that $j < t$, observing both $x \succ_C^{rs} y$ and $y \succ_C^{t'} x$ would imply, given the definition of $\succ_C^{rs}$and $\succ_C^{t'}$and the transitivity of $\succsim_1$, that both $x \succ_1 y$ and $y \succ_1 x$ holds, which is also a contradiction. Since $r < t \leq t'$, one has that $a_v = C(v, A(v))$ if and only if $a_v \succsim_2 a'_v$ for all $a'_v \in A(v)$ and all v such that $v \geq t$. But then, the assumed existence of w and $z \in X$ for which one has $w \succsim_C^{uv} z$ and $z \succ_C^{\tau} w$ for some periods u, v and τ such that $t' \leq u \leq v < \tau$ leads to the conclusion that both $w \succ_2 z$ and $z \succ_2 w$ holds, which is a contradiction. Hence a chronological choice function that results from one change in preferences satisfies Axiom 2.

In order to prove the converse implication, consider a chronological choice function that satisfies Axiom 2. If there exists no r, s, $t \in \mathcal{T}$ such that $1 \leq r < s < t$ for which one has $x \succ_C^{rs} y$ and $y \succ_C^{t} x$ for some distinct $x, y \in X$, then this means that the chronological choice function satisfies Axiom 1. In that case, set $\succsim_1 = \succsim$ where $\succsim$ is the linear ordering whose existence was established in Theorem 1 and let $\succsim_2$ be any linear

ordering whatsoever. As shown in Theorem 1, the linear ordering $\succsim$ will rationalize the behavior of the chronological choice function C from 1 up to T.

Suppose now that there exists some $r, s, t \in \mathcal{T}$ such that $1 \leq r < s < t$ for which one has $x \succ_C^{rs} y$ and $y \succ_C^{t} x$ for some distinct $x, y \in X$. Define then $\widehat{t}$ to be the smallest such t. By Axiom 2, the chronological choice function satisfies Axiom 1 on the time horizon $\{1, ..., \widehat{t} - 1\}$ and it also satisfies Axiom 1 on the (non-empty) time horizon $\{\widehat{t}, ..., T\}$. The result then follows from applying Theorem 1 to the time horizons $\{1, ..., \widehat{t} - 1\}$ and $\{\widehat{t}, ..., T\}$ sequentially. □

Theorem 2 hence provides an easy way to test if the behavior of an agent is consistent with changing preferences at most once, and choosing at each period as per the preference of this period. Remark that Theorem 2 easily extends to the case with more than one change in preferences. This could be done by just rewriting Axiom 2 for k-changes, and applying it for $k + 1$ partitions of the time horizon in the proof. Of course, if the number of changes in preferences is equal to the number of periods, then any choice behavior can be rationalized (see Kalai et al. 2002 for a similar observation in the timeless setting).

The changing preferences choice model examined in this section is somewhat different from the *revealed preference theory of changing tastes* analyzed by Gul and Pesendorfer (2005). The authors characterize a model of consistent planning, that rationalize changing tastes due to temptation and self-control. On the one hand, the changing preferences model examined in this section is more general than theirs since it allows for any source of change in preferences. On the other hand, and contrary to Gul and Pesendorfer (2005), our model is silent on the effect of current choices on the shape of the future menus of available alternatives. Indeed, in our approach, the chronology of choices is exogenously given and it is not affected by the choices made by the agent. It would be interesting to allow the chronology of choices to be affected by the chronological choice function.

3.4.2 Learning by Trial and Error

We now consider the possibility for a decision maker to behave as if she was forming her preference between two alternatives only after the two alternatives have been previously "tried" at least once. This choice model is consistent with "rational behavior", but accommodates some learning that may lead to some initial "contradictions" in choices. Hence, we require the decision maker to be consistent in her choices in the sense of Axiom 1 *only* when those choices concern alternatives that have been tried at least once in the past.

To illustrate the model we have in mind, we find again useful to consider the following two examples.

Example 3. *Let $\mathcal{T} = \{1, 2, 3, 4\}$ and consider again a chronology of (gastronomic) menus:*

$A(1) = \{chicken, dahl\}$, $A(2) = \{beef, dahl\}$, $A(3) = \{beef, chicken, dahl\}$ and $A(4) = \{beef, chicken\}$.

The chronological choice function C defined by $C(1, A(1)) = chicken$, $C(2, A(2)) = beef$, $C(3, A(3)) = chicken$, and $C(4, A(4)) = beef$ is not consistent with a learning by trial and error model. In the first period, without any information about her preferences for food, the decision maker goes for chicken and experienced the taste. In the second period she goes for beef and tries out its taste. In the third period, where she has the choice between chicken, beef, and dahl, she reveals a preference for chicken over beef. Given that she knows the tastes (because she has tried both in the past), the choice in the third period reveals a "definite" preference for chicken over beef. But then the choice at the fourth period - beef over chicken - is inconsistent with this preference.

Example 4. *Let $\mathcal{T} = \{1, 2, 3, 4\}$ and consider the following chronology:*

$A(1) = \{beef, chicken, dahl\}$, $A(2) = \{beef, chicken\}$, $A(3) = \{chicken, dahl\}$ and $A(4) = \{beef, dahl\}$.

The chronological choice function C defined by $C(1, A(1)) = chicken$, $C(2, A(2)) = beef$, $C(3, A(3)) = chicken$, and $C(4, A(4)) = beef$ describes a behavior consistent with a

learning by trial and error model. Indeed, albeit one observes revealed preferences "inconsistencies" (in the traditional sense) between the choices made at the two first periods, these inconsistencies may be interpreted as the results of trial and error. Indeed, the decision maker may be trying out chicken in the first period and trying out beef at the second. After these trials, the decision maker reveals a "definite" preference for chicken over beef (in period 3), and in this example she is consistent with it in the following period.

We emphasize the crucial importance of introducing time for characterizing a behavior resulting from preference formation by trial and error. Indeed, the only difference between the two examples is the time order at which the menus - identical in both examples - appear. Hence, without time, the two choice behaviors could not be distinguished and, as a result, it would not be possible to identify those violations of standard revealed preference that are compatible with a process of preference formation through trial and error and those violations that are not so.

In order to characterize the behavioral implications of this model, we find it useful to define the following "revealed definitely preferred" binary relation.

Definition 5. *For any period $t \in \mathcal{T}$ and some x and $y \in X$, we say that x* ***is directly revealed definitely preferred to*** *y* ***at period*** *t, denoted $x \succsim_C^{Dt} y$, if and only if there are periods r, s, and t in $\mathcal{T}$ satisfying $r < t$ and $s < t$ such that $x = C(r, A(r)) = C(t, A(t))$, $y \in A(t)$ and $y = C(s, A(s))$.*

In words, the chronological choice function directly reveals a *definite preference* for x over y at period t (with x and y distinct) whenever it reveals (by choice) a preference for x over y at a period t that follows periods where x and y have been tried. Since both x and y have been tried before t, one can interpret the choice of x over y in period t as revealing a "definite" preference between the two alternatives.

Given this "direct revealed definite preference at period t" relation, one defines the revealed definite preference relation over a sequence of periods going from r up to s as follows.

Definition 6. *For any periods r and s such that $r \leq s$ and some x and $y \in X$, we say that x* ***is indirectly revealed definitely preferred to*** *y* ***between periods*** *r* ***and*** *s, denoted $x \succsim_C^{Drs} y$, if and only if there is a sequence $\{t_j\}_{j=1}^k$ of k time periods in the set $\{r, r+1, ..., s-1, s\}$, not necessarily ordered by time, for which one has:*
(i) $x = C(t_1, A(t_1))$,
(ii) $C(t_j, A(t_j)) \succsim_C^{Dt_j} C(t_{j+1}, A(t_{j+1}))$ for all $j = 1, ..., k-1$, and
(iii) $y \in A(t_k)$ and $y = C(t_h, A(t_h))$ for some $t_h < r$.

The following axiom will be shown to be necessary and sufficient for a chronological choice function to be rationalized by a preference formation procedure through trial and error.

Axiom 3. *For any periods r, s and t such that $r \leq s < t$ and some distinct x and $y \in X$, one cannot have $x \succsim_C^{Drs} y$ and $y \succ_C^{Dt} x$.*

In plain English, this axiom says that we should never observe a violation of Axiom 1 for two alternatives that have been previously chosen at least once in the past. We now define what is meant by a chronological choice behavior to result from the maximization of a preference formed by trial and error.

Definition 7. *A chronological choice function C results from the maximization of a preference formed by trial and error if there exists a linear ordering $\succsim$ on X such that, for all $t \in \mathcal{T}$, either $a_t = C(t, A(t))$ if and only if $a_t \succsim a'_t$ for all $a'_t \in A(t)$ for which $a'_t = C(s, A(s))$ for some $s < t$, or there is no $s' \in \mathcal{T}$ such that $C(t, A(t)) = C(s', A(s'))$ and $s' < t$.*

That is, a chronological choice behavior results from the maximization of a preference formed by trial and error if there exists a linear preference such that the choice

made by the decision maker at every period is either the "best" option for that preference among all alternatives that have been previously tried or, if this is not the case, it is because the chosen option has never been tried before. Note that the "best" option for that preference among all alternatives that have been previously tried is not necessarily the maximal option for that preference. It is possible that the maximal option has itself never been tried before. Of course, in this case, the decision maker "does not know" yet that this option is maximal.

It is easy to show that Axiom 3 is a necessary and sufficient condition for a chronological choice function to result from the maximization of a preference formed by a trial and error process.

Theorem 3. *A chronological choice function C satisfies Axiom 3 if and only if it results from the maximization of a preference formed by trial and error.*

Proof. For the "if" part of the theorem, assume by contradiction that a chronological choice function C results from the maximization of a preference formed by trial and error but that it violates Axiom 3 Hence, there are periods r, s and t in $\mathcal{T}$ such that $r \leq s < t$ and some distinct x and $y \in X$, for which one have $x \succsim_C^{Drs} y$ and $y \succ_C^{Dt} x$. By definition of $x \succsim_C^{Drs} y$, there is a sequence $\{t_j\}_{j=1}^{k}$ of k time periods in the set $\{r, r+1, ..., s-1, s\}$ such that $x = C(t_1, A(t_1))$, $C(t_j, A(t_j)) \succsim_C^{Dt_j} C(t_{j+1}, A(t_{j+1}))$ for all $j = 1, ..., k-1$, and $y \in A(t_k)$ and $y = C(t_h, A(t_h))$ for some $t_h < r$. By definition of $C(t_j, A(t_j)) \succsim_C^{Dt_j} C(t_{j+1}, A(t_{j+1}))$ for all $j = 1, ..., k-1$, there are, for any such j, periods r_j, and s_j in $\mathcal{T}$ satisfying $r_j < t_j$ and $s_j < t_j$ such that $C(r_j, A(r)) = C(t_j, A(t_j))$, $C(t_{j+1}, A(t_{j+1})) \in A(t_j)$ and $C(t_{j+1}, A(t_{j+1})) = C(s_j, A(s_j))$. Since C results from the maximization of a preference formed by trial and error, there exists a linear ordering $\succsim$ on X such that $x = C(t_1, A(t_1)) \succsim C(t_2, A(t_2)) \succsim ... \succsim C(t_k, A(t_k))$ for all $j = 1, ..., k-1$. Since $y \in A(t_k)$ holds and C results from the maximization of a preference formed by trial and error, one has $C(t_k, A(t_k)) \succsim y$. By the transitivity and the linearity of $\succsim$ (as x and y are distinct) one has $x \succ y$. But then, assuming $y \succ_C^{Dt} x$ for $t > s \geq r$ implies,

under the assumption that C results from the maximization of a preference formed by trial and error, that $y \succ x$, which is a contradiction.

To prove the other direction of the implication, consider a chronological choice function C that satisfies Axiom 3 and define the following "definite revealed preference" relation $\succsim_C^D$:

$$\begin{aligned} x \ \succsim_C^D y \iff & \exists r, s,\ t \in \mathcal{T} \text{ satisfying } r < t \text{ and } s < t \text{ such that:} \\ x \ = \ & C(t, A(t)),\ y \in A(t),\ x = C(r, A(r)) \text{ and } y = C(s, A(s)) \end{aligned}$$

Notice that this binary relation can be empty. This would happen, for example, for a chronology in which the same menu is available at every period and a chronological choice function that chooses the same alternative from that same menu at every period. In such a trivial case, the decision maker would never experience anything other than this chosen option, and there would therefore be no pair of alternatives between which the binary relation $\succsim_C^D$ would hold. That is, the decision maker would never be given the opportunity to express a "definite preference". In such a case, the choice behavior can be (trivially) rationalized by any linear ordering $\succsim$ whatsoever. Indeed, take any linear ordering $\succsim$ and consider any period t for which $x \succsim a_t$ for some $x \in A(t)$ and all $a_t \in A(t)$ but for which $x \neq C(t, A(t))$. There may not be any such t, in which case the linear ordering $\succsim$ rationalizes the choice behavior in the usual sense. If however such a t exists, we then know from the emptiness of the binary relation $\succsim_C^D$that either $C(t, A(t)) \neq C(s, A(s))$ for all $s < t$ or $x \neq C(s, A(s))$ for all $s < t$. Hence the chronological choice behavior is trivially rationalized as resulting from the maximization of preference formed by trial and error when $\succsim_C^D$ is empty. If $\succsim_C^D$ is not empty, one can

define its transitive closure $\widehat{\succsim}_C^D$ by:

$$\begin{aligned} x \,\widehat{\succsim}_C^D\, y \iff & \exists\, \{x_j\}_{j=0}^{l} \text{ for some } l \geq 1 \text{ such that:} \\ x_0 = & \; x, \\ x_l = & \; y \text{ and,} \\ x_j \succsim & \; {}_C^D x_{j+1} \text{ for all } j = 0, ..., l-1 \end{aligned}$$

Let us now show that the (transitive) binary relation $\widehat{\succsim}_C^D$ is also antisymmetric. By contradiction, suppose there are two distinct x and $y \in X$ such that $x \,\widehat{\succsim}_C^D\, y$ and $y \,\widehat{\succsim}_C^D\, x$. By definition of $\widehat{\succsim}_C^D$ and $\succsim_C^D$, there are two sequences of triples of periods $\{r_j, s_j, t_j\}_{j=0}^{l}$ and $\{r'_j, s'_j, t'_j\}_{j=0}^{l'}$ (for some l and $l' \geq 1$) satisfying, for every j, $r_j < t_j$, $s_j < t_j$, $r'_j < t'_j$ and $s'_j < t'_j$ for which one has:

$$x_j = C(r_j, A(r_j)) = C(t_j, A(t_j)),\ x_{j+1} = C(s_j, A(s_j)) \text{ and } x_{j+1} \in C(t_j, A(t_j))$$

as well as :

$$x'_j = C(r'_j, A(r'_j)) = C(t'_j, A(t'_j)),\ x'_{j+1} = C(s'_j, A(s'_j)) \text{ and } x'_{j+1} \in C(t'_j, A(t'_j))$$

for two sequences of alternatives $\{x_j\}_{j=0}^{l}$ and $\{x'_j\}_{j=0}^{l'}$ satisfying $x_j \in X$, $x'_j \in X$ for all j, $x_0 = x'_{l'} = x$ and $x_l = x'_0 = y$. This generates a cycle of revealed definite preference connecting alternatives in X that can be initiated at every period of the sets of periods $\{t_j\}_{j=0}^{l}$ and $\{t'_j\}_{j=0}^{l'}$ defined above. In particular, one can take the maximal (with respect to the natural ordering of time) of this period, and apply the reasoning of the proof of Theorem 1 to obtain the required violation of Axiom 3. Since $\widehat{\succsim}_C^D$ is antisymmetric and transitive, it can be extended to a linear ordering $\succsim$ using Spilrajn extension theorem. Let us now prove that the chronological choice function C results from the maximization of $\succsim$ formed by a trial and error process. Consider any $t \in \mathcal{T}$. Either $C(t, A(t)) \succsim a_t$ for all $a_t \in A(t)$ or $\exists x \in A(t)$ such that $x \neq C(t, A(t))$ and $x \succsim C(t, A(t))$. In the first case, $\succsim$ rationalizes the choice made in the choice problem at t and there is nothing to prove. In the second case, take without loss of generality the alternative $x \in A(t)$

to be such that $x \succsim a_t$ for all $a_t \in A(t)$ By assumption $x \neq C(t, A(t))$. Suppose that, contrary to the requirement that the chronological choice function C results from the maximization of $\succsim$ formed by a trial and error process, it is neither the case that $C(t, A(t)) \neq C(s, A(s))$ for all $s < t$ nor $x \neq C(s, A(s))$ for all $s < t$. This means that there exists a period $r < t$ such that $x = C(r, A(r))$ and there exists a period $s < t$ such that $C(t, A(t)) = C(s, A(s))$. It then follows from the definition of $\succsim_C^D$ that $C(t, A(t)) \succsim_C^D x$ and, since $\succsim$ is an extension of $\succsim_C^D$, that $C(t, A(t)) \succsim x$. This means that we have both $x \succsim C(t, A(t))$ and $C(t, A(t)) \succsim x$, a contradiction of $\succsim$ being antisymmetric. □

The trial and error method of learning seems quite plausible as a way to discover one's preference. For example, children learn that they prefer apples over bananas by trying both at different times, and by "discovering" that they indeed prefer apples to bananas. Once this discovery is made - and provided that no subsequent change in preferences take place - children will stick to this preference and never choose a banana when an apple is also available. In economics, the trial and error method of learning has been studied for organizational learning such as within-firm experimentation (see e.g Nelson 2008 and Callander 2011). Young (2009) has also examined a trial and error learning rule in a game theoretical environment while Cooke (2016) and Piermont et al. (2016) have examined models of preference formation through experimentation over uncertain prospects. To the best of our knowledge, Theorem 3 is the only available characterization of a learning process by trial and error over an abstract set of alternatives.

The learning model characterized in Theorem 3 provides a simple rationalization of "inconsistent" behavior at the beginning of some sequence of choices. For example, one could think that a trial and error method could be used in order to form a preference over different alternatives in an experimental design with which subjects are not familiar with. Clearly, many other models of learning could be characterized within this framework in order to explain initial inconsistencies in behavior. Similarly, initial inconsistencies can be rationalized by a model that allows for initial mistakes.

An open question is if these alternative explanations would be observationally equivalent to a model of learning by trial and error. But in case they would have different empirical implications, one could use this framework to test different explanations for the same observable behavior.

3.4.3 Choice with Inertia Bias

Another theory of choice behavior whose behavioral implications can be characterized by means of a chronological choice function is that of a decision maker who has a bias towards her (immediately) last choice. One interpretation of such behavior is that the decision maker has inertia in her preferences and sees $C(t-1, A(t-1))$ as a default option when making a choice at period t. Another interpretation is that the decision maker has an "imperfect recall" of the choices that took place earlier in the time horizon, and takes the previous choice as a *status-quo* option. In our setting, this means that for each choice problem, the decision maker either chooses the best option according to a time-invariant preference *or* chooses the option that she has chosen in the previous choice problem.

The following examples illustrate the behavioral implications of this theory of choice.

Example 5. *Let $\mathcal{T} = \{1, 2, 3\}$ and consider the following chronology:*
$A(1) = \{chicken, beef\}$, $A_2 = \{beef, dahl\}$, $A_3 = \{chicken, beef, dahl\}$.
The chronological choice function C defined by $C(1, A(1)) = C(2, A(2)) = beef$ and $C(3, A(3)) = chicken$ is not consistent with a model of choice with inertia bias. Note that the decision maker has chosen to eat chicken in the last period, while her immediately preceding choice - beef - was available. Hence our decision maker has "broken" her inertia by choosing something else than her last choice. If this "break" is motivated by a desire to obtain a preferable alternative for a well-defined time-invariant preference, as assumed in this model, then the ranking of alternatives provided by this preference must be the same at every period at which the preference is expressed. When can we be (more)

confident that such preference is expressed? When the choice made at some period is different from the choice made at the immediately preceding period or, trivially, at the beginning of the history (when there is no past and, therefore, no source of inertia). Yet, here, in the first period, the decision maker has revealed a preference for beef over chicken, that is inconsistent with the "active"[3] preference (as opposed to inertia) revealed by her choice in the last period.

Example 6. *Let again* $\mathcal{T} = \{1, 2, 3\}$ *and consider the same chronology as before:* $A(1) = \{chicken, beef\}$, $A_2 = \{beef, dahl\}$, $A_3 = \{chicken, beef, dahl\}$. *The chronological choice function* C *defined by* $C(1, A(1)) = chicken$, $C(2, A(2)) = C(3, A(3)) = beef$, *is consistent with a model of choice with inertia bias. Notice that the chronological choice behavior violates Axiom 1. Indeed, the preference for beef over chicken revealed, in the traditional sense, in the last choice period is inconsistent with the preference for chicken over beef revealed in the first period. However, the alternative chosen in the third period is also the one that was chosen in the second period. Hence, the choice of the third period cannot be interpreted as revealing an active preference. It may also be the result of an inertia bias.*

Again, these two examples could not be distinguished without the introduction of time. As in the previous subsection, we find convenient to redefine the revealed preference relations in a way that is suitable for identifying an inertia bias explanation of choices. As discussed in the two examples, when the default option is present and chosen, the observer of the choice does not know if it reveals an active preference for the chosen option over the non-chosen ones or if it results from an inertia bias. We accordingly define the notions of direct and indirect active preferences as follows.

[3] We avoid terms such as "authentic" or "true" preference since we just wish to observationally distinguish between choices that may result from a time-invariant preference and choices that may result from an inertia bias. This exercise may be useful, for instance, when there is a (strong) reason to associate such inertia bias to a mistake. However, we do not wish to interpret *a-priori* a potential underlying stable preference as the "true" and/or normatively relevant preference of the decision maker (see Chapter 1 and Infante et al. 2016 for the potential difficulties of such interpretation).

Definition 8. *For any period t and some x and $y \in X$, we say that x* ***is directly revealed actively preferred to*** *y* ***at period*** *t, denoted $x \succsim_C^{At} y$, if and only if $x = C(t, A(t))$, $y \in A(t)$ and either $t = 1$ or $x \neq C(t-1, A(t-1))$.*

Definition 9. *For any periods r and s such that $r \leq s$ and some x and $y \in X$, we say that x* ***is indirectly revealed actively preferred to*** *y* ***between periods*** *r* ***and*** *s, denoted $x \succsim_C^{Ars} y$, if and only if there is a sequence $\{t_j\}_{j=1}^k$ of k time periods in the set $\{r, r+1, ..., s-1, s\}$, not necessarily ordered by time, for which one has:*
(i) $x = C(t_1, A(t_1))$
(ii) $C(t_j, A(t_j)) \succsim_C^{At_j} C(t_{j+1}, A(t_{j+1}))$ for all $j = 1, ..., k-1$, and
(iii) $y \in A(t_k)$.

We now non-surprisingly formulate the axiom which characterizes a behavior described by a chronological choice function which results from inertia bias.

Axiom 4. *For any periods r, s and t such that $r \leq s < t$ and some distinct x and $y \in X$, one cannot have $x \succsim_C^{Ars} y$ and $y \succ_C^{At} x$.*

We can also define what is meant by a chronological choice behavior to result from a choice model with inertia bias.

Definition 10. *A chronological choice function results from choice with inertia bias if there exists a linear ordering $\succsim$ on X such that, for all $t \in \mathcal{T}$, either $x = C(t, A(t))$ if and only if $x \succsim a_t$ for all $a_t \in A(t)$ or $t > 1$ and $C(t, A(t)) = C(t-1, A(t-1))$.*

Then, one can establish the following:

Theorem 4. *A chronological choice function C satisfies Axiom 4 if and only it it results from choice with inertia bias.*

Proof. As the argument is very similar to those of Theorems 2 and 3, we only sketch the proof, and leave to the reader the task of verifying that a chronological choice function that results from choice with inertia bias satisfies Axiom 4. As for the converse

implication, consider a chronological choice function C that satisfies Axiom 4 and define the following "active revealed preference" relation $\succsim_C^A$:

$x \succsim_C^A y \iff \exists t \in \mathcal{T}$ s. t. $x = C(t, A(t)),\ y \in A(t)$ and either $C(t-1, A(t-1)) \neq C(t, A(t))$ or $t = 1$:

This binary relation is not empty [because $C(1, A(1)) \succsim_C^G y$ for every $y \in A(1)$]. One can then define its transitive closure $\widehat{\succsim}_C^A$ by:

$$\begin{aligned} x \,\widehat{\succsim}_C^A\, y \iff & \exists \{t_j\}_{j=0}^{l} \text{ for some } l \geq 1, \text{ with } t_j \in \mathcal{T} \text{ for all } j \text{ such that:} \\ x = & C(t_0, A(t_0)), \\ y \in & A(t_l) \text{ and,} \\ C(t_{j+1}, A(t_{j+1}) \in & A(t_j) \text{ for all } j = 0, ..., l-1 \text{ and} \\ \text{either } C(t_j - 1, A(t_j - 1)) \neq & C(t_j, A(t_j)) \text{ or } t_j = 1 \text{ for all } j = 0, ..., l \end{aligned}$$

Let us now show that the (transitive) binary relation $\widehat{\succsim}_C^A$ is antisymmetric. By contradiction, suppose there are two distinct x and $y \in X$ such that $x \,\widehat{\succsim}_C^A\, y$ and $y \,\widehat{\succsim}_C^A\, x$. Using an analogous reasoning as in Theorems 2 and 3, this would generate a cycle of revealed preferences that would be inconsistent with Axiom 4. Hence $\widehat{\succsim}^A$ must be antisymmetric and transitive. It can therefore be extended to a linear ordering $\succsim$ using Spilrajn extension theorem just as before. We just need to prove that C is such that, for every period $t \in \mathcal{T}$, either $x = C(t, A(t))$ if and only if $x \succsim a_t$ for all $a_t \in A(t)$ or $t > 1$ and $C(t, A(t)) = C(t-1, A(t-1))$. Consider first $t = 1$. By definition of $\succsim_C^A$, one has $C(1, A(1)) \succsim_C^A a_1$ for all $a_1 \in A(1)$ and, since $\succsim$ is an extension of $\succsim_C^A$, one has $C(1, A(1)) \succsim a_1$ for all $a_1 \in A(1)$ as well. Moreover the antisymmetry of $\succsim$ prevents any alternative x of $A(1)$ distinct from $C(1, A(1))$ to be such that $x \succsim a_1$ for all $a_1 \in A(1)$. Hence one has $x = C(1, A(1)) \iff x \succsim a_1$ for all $a_1 \in A(1)$. Consider now any period $t > 1$. Assume $x = C(t, A(t))$. Either $C(t, A(t)) = C(t-1, A(t-1))$ or $C(t, A(t)) \neq C(t-1, A(t-1))$. There is nothing to be proved in the first case. In the second case, one has $C(t, A(t)) \succsim_C^A a_t$ for all a_t by definition of $\succsim_C^A$ and $C(t, A(t)) \succsim a_t$ for all a_t

in $A(t)$ by definition of the linear ordering $\succsim$ to be an extension of $\widehat{\succsim}_C^A$. Since, as just established, $C(t, A(t)) \succsim a_t$ for all a_t in $A(t)$ and $\succsim$ is linear, there cannot be a $z \in A(t)$ distinct from $C(t, A(t))$ such that $z \succsim a_t$ for all a_t in $A(t)$. Hence one has $x = C(t, A(t))$ if and only if $x \succsim a_t$ for all a_t in $A(t)$, as required. □

The model of choice characterized by Theorem 4 can be connected with the numerous models of choice with reference-dependent preferences and *status-quo* bias discussed in the literature (e.g. Tversky and Kahneman 1991) and characterized axiomatically (see e.g. Bossert and Sprumont 2003; Masatlioglu and Ok 2005). However, all previous characterizations that we are aware of have considered a default alternative that is fixed and exogenous. We depart here from the literature by characterizing a model with an endogenous process for the formation of a *status-quo* bias that could be interpreted a coming from an "imprisonment in the habits" phenomenon. We have given two potential interpretations for this phenomenon, either as inertia or imperfect recall of (remote) past choices. For example, inertia in preferences has been documented in management and economics literature (see e.g. Dubé et al. 2010 for inertia in brand choice). Finally, it can be easily verified that the choice model characterized by Theorem 4 is not observationally equivalent to those involving exogenous *status-quo* bias and could therefore be tested against these models in experimental contexts.

3.5 Concluding Remarks

In this chapter, we have argued in favor of explicitly introducing time in the description of choice behavior provided by a choice function. We have used our setting to characterize the behavioral implications of three alternative theories of choice. We end this chapter with a brief discussion of some potential limitations of the approach and possible extensions.

First, we have limited our attention to choice functions, but one could wish to extend this analysis for choice correspondences. As noted in Section 3.3, Suzumura

(1976) has shown that SARP is a necessary and sufficient condition for a choice correspondence to be rationalized by a complete and acyclical binary relation. But a choice correspondence could be used to find similar characterizations for other one-rationale or multiple-rationale choice theories. A related - but distinct - possibility would be to use a chronological choice function to induce a standard timeless choice correspondence as done in Bernheim and Rangel (2007; 2009) or Salant and Rubinstein (2008). In our framework, if the menus $A(t)$ and $A(t')$ offered to the decision maker at distinct time periods t and t' where the same (and, say, equal to the set A) and if $C(t, A(t)) \neq C(t'A(t'))$, then the timeless choice correspondence C_c induced by the chronological choice function C would yield $C_c(A) = \{C(t, A(t)), C(t'A(t'))\}$.

Second, we have taken time to be a sequence of "unrelated" choice periods (except with respect to their order), but one could wish to establish further links between "similar" periods over time. For example, in some applications the day of the week may influence behavior, and some consistency could be required for the same day (say Mondays) when observed more than once. This could, at least in principle, allow an observer to impose consistency between the choices made in different sequences, such as one observed sequence (say one week) and a non-observed future sequence (say the week after). One could, in such settings, explore explanations for preference reversals or cognitive biases that "return" over and over and affect repeatedly the choice behavior of an agent. Another possibility would be to record the frequency of choices for the same menu over time. Resorting to this information, and provided that enough repetition was observed for the same menu, an observer could try to predict behavior at the limit.

Third, we have limited our attention to deterministic choice, but one could wish to study stochastic choice in this setting. Time, if taken to be not only a sequence of choices but also a repetition of events (as in the calendar example above), seems to be correlated with repeated states of the world. Then, it would be interesting - though potentially complex - to use our framework to study, for instance, random preferences models according to which agents' preferences change stochastically (e.g.

Becker et al. 1963; Barberà and Pattanaik 1986; McFadden and Richter 1990; Loomes and Sugden 1995; Gul and Pesendorfer 2006; Apesteguia et al. 2017).

Fourth, we note that while the identification of the behavioral implications of the chronological choice functions characterized in this chapter could lead to interesting empirical or experimental applications, these implications are formulated in terms of indirect revealed preference relations. While this is quite standard in the choice theoretic literature, we emphasize that the empirical tests of such revealed preference axioms may be computationally demanding if the universe of alternatives is large.

Fifth, a characteristic shared by the three time-dependent choice models that we characterize is that they do not have a "strong structure", i.e., the implications that they impose upon choice behavior are not very restrictive. This means that these theories can rationalize very different patterns of behavior. This can be either seen as a strength or a weakness of these models. On the one hand, this may lead to problems related to indistinguishability discussed in Chapter 2. By observing a given choice behavior that is consistent with one of these models, we cannot be "too" confident that this is indeed the decision making model that underlays that behavior. On the other hand, these theories have a wide scope and are not too demanding in terms of rationality. Nonetheless, they establish meaningful restrictions upon choice behavior.

Finally, we find worth pointing out the ease by which the characterization of the choice behavior exhibited in the three examples examined herein was obtained. Hence the simple fact of introducing time in the description of choice behavior seems to have the significant payoff of alleviating what Rubinstein (2012) calls "the burden on researchers" of finding the observable properties of the behavioral decision making models that they are interested in.

Bibliography

Afriat, S. (1967). The construction of utility functions from expenditure data. *International Economic Review*, 8(1):67–77.

Apesteguia, J. and Ballester, M. A. (2013). Choice by sequential procedures. *Games and Economic Behavior*, 77(1):90–99.

Apesteguia, J., Ballester, M. A., and Lu, J. (2017). Single crossing random utility models. *Econometrica*, 85(2):661–74.

Arrow, K. J. (1959). Rational choice functions and orderings. *Economica*, 26:121–27.

Barberà, S. and Neme, A. (2016). Ordinal relative satisficing behavior: Theory and experiments. *Barcelona GSE Working Paper Series 790*.

Barberà, S. and Pattanaik, P. K. (1986). Falmagne and the rationalizability of stochastic choices in terms of random orderings. *Econometrica*, 54(3):707–15.

Becker, G., DeGroot, M., and Marschak, J. (1963). Stochastic models of choice behavior. *Behavioral Science*, 8:41–55.

Bernheim, B. D. and Rangel, A. (2007). Toward choice-theoretic foundations for behavioral welfare economics. *The American Economic Review: Papers and Proceedings*, 97(2):464–70.

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

Bossert, W. and Sprumont, Y. (2003). Non-deteriorating choice. *Mathematical Social Sciences*, 45:131–42.

Bossert, W., Sprumont, Y., and Suzumura, K. (2005). Consistent rationalizability. *Economica*, 72:185–200.

Bossert, W., Sprumont, Y., and Suzumura, K. (2006). Rationalizability of choice functions on general domains without full transitivity. *Social Choice and Welfare*, 27:435–58.

Callander, S. (2011). Searching and learning by trial and error. *The American Economic Review*, 101:2277–308.

Cerigioni, F. (2016). Dual decision processes: Retrieving preferences when some choice are intuitive. *Barcelona GSE Working Paper Series 924*.

Chernoff, H. (1954). Rational selection of decision functions. *Econometrica*, 30:918–925.

Cooke, K. (2016). Preference discovery and experimentation. *Theoretical Economics (Forthcoming)*.

Dubé, J.-P., Hitsch, G. J., and Rossi, P. E. (2010). State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, 41(3):417–45.

Fehr, E. and Hoff, K. (2011). Introduction: Tastes, castes and culture: The influence of society on preferences. *The Economic Journal*, 211:396–412.

Fudenberg, D. (2006). Advancing beyond advances in behavioral economics. *Journal of Economic Literature*, 44:694–711.

Gul, F. and Pesendorfer, W. (2005). The revealed preference theory of changing tastes. *The Review of Economic Studies*, 72(2):429–48.

Gul, F. and Pesendorfer, W. (2006). Random expected utility. *Econometrica*, 74(1):121–46.

Hansson, B. (1968). Choice structures and preference relations. *Synthese*, 18:443–58.

Hoff, K. and Stiglitz, J. E. (2016). Striving for balance in economics: Towards a theory of the social determination of behavior. *Journal of Economic Behavior and Organization*, 126:25–57.

Houthakker, H. (1950). Revealed preference and the utility function. *Economica*, 17:159–74.

Infante, G., Lecouteux, G., and Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, 23(1):1–25.

Kalai, G., Rubinstein, A., and Spiegler, R. (2002). Rationalizing choice function by multiple rationales. *Econometrica*, 70(6):2481–88.

Loomes, G. and Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, 39:641–48.

Manzini, P. and Mariotti, M. (2007). Sequentially rationalizable choice. *The American Economic Review*, 97(5):1824–39.

Manzini, P. and Mariotti, M. (2012). Choice by lexicographic semiorders. *Theoretical Economics*, 7:1–23.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed attention. *The American Economic Review*, 102(5):2183–205.

Masatlioglu, Y. and Ok, E. (2005). Rational choice with status quo bias. *Journal of Economic Theory*, 111(1):1–29.

McFadden, D. and Richter, M. K. (1990). Stochastic rationality and revealed stochastic preference. In Chipman, J. S., McFadden, D., and Richter, M. K., editors, *Preferences, Uncertainty, and Optimality: Essays in Honor of Leo Hurwicz*, pages 163–186. Westview Press: Boulder, CO, 161-186., Boulder, Colorado.

Nelson, R. R. (2008). Bounded rationality, cognitive maps, and trial and error learning. *Journal of Economic Behavior and Organization*, 67:78–89.

Piermont, E., Takeoka, N., and Teper, R. (2016). Learning the krepsian state: Exploration through consumption. *Games and Economic Behavior*, 100:69–94.

Richter, M. K. (1966). Revealed preference theory. *Econometrica*, 34:635–45.

Richter, M. K. (1971). Rational choice. In Chipman, J., Hurwicz, L., Richter, M., and Sonnenschein, H., editors, *Preferences, Utility, and Demand*, pages 29–58. Harcourt Brace Jovanovich, New York.

Rubinstein, A. (2012). *Lecture Notes in Microeconomic Theory*. Princeton University Press, Princeton, N.J., 2nd edition edition.

Salant, Y. and Rubinstein, A. (2008). (a, f): Choice with frames. *The Review of Economic Studies*, 75(4):1287–96.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71.

Sen, A. K. (1971). Choice functions and revealed preferences. *Review of Economic Studies*, 38:307–17.

Sprumont, Y. (2000). On the testable implications of collective choice theories. *Journal of Economic Theory*, 93:205–232.

Suzumura, K. (1976). Rational choice and revealed preference. *Review of Economic Studies*, 43:149–58.

Suzumura, K. (1977). Houthakker's axiom in the theory of rational choice. *Journal of Economic Theory*, 14:284–90.

Suzumura, K. (1983). *Rational Choice, Collective Decisions and Social Welfare*. Cambridge University Press., New York.

Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 107(4):1039–1061.

Varian, H. (2006). Revealed preferences. In Ramrattan, M. S. L. and Gottesman, A., editors, *Samuelsonian Economics and the Twenty First Century*, pages 99–115. Oxford University Press, Oxford.

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica*, 50(4):945–72.

Young, H. P. (2009). Learning by trial and error. *Games and Economic Behavior*, 65:626–43.

Chapter 4

Conflicted Voters: A Spatial Voting Model with Multiple Party Identifications*

We develop a unified spatial model of turnout and voting behaviors in which citizens can identify with one or two parties. We show the existence of a *conflicted voter's curse*: If there is no position that reconciles the ideological views of both parties, it is always rational for citizens that identify with two parties to abstain even if they are a majority. In a two-candidate electoral competition, the conflicted voter's curse implies that candidates converge to the center of the political domain if and only if conflicted voters are pivotal and the parties have shared ideological views. Otherwise, we show that candidates may converge or diverge depending upon the degree of party polarization and whether candidates care about ideology or not. Our analysis illustrates how the behavior of conflicted voters may be relevant for electoral outcomes.

Keywords: Spatial Voting; Party identification; Conflicted voters; Electoral competition; Party polarization.

*This chapter is adapted with some liberty from a joint work with Sacha Bourgeois-Gironde, with the same title, forthcoming in the *Journal of Economic Behavior and Organization*.

> "Any party which is both responsible and reliable will probably have an ideology which is relatively coherent and immobile. In other words, its ideology will not be internally contradictory but will be at least loosely integrated around some social *Weltanschauung*. And the party will not radically shift its policies and doctrines overnight, but will only slowly change their nature."
>
> (Downs 1957, 109)

4.1 Introduction

After the seminal contributions of Hotelling (1929), Downs (1957), and Black (1958), the spatial theory of voting became a cornerstone to the study of elections. The aim of this literature is to study the electoral outcomes that emerge from the interaction between two economic agents: candidates (key actors) and citizens (fixed role). Citizens are assumed to have exogenous preferences over a uni or multidimensional ideological space, and candidates compete for election by adopting positions in this space. Amid the numerous contributions to this literature the term "candidate" is often interchanged with "party", and most papers abstract from any distinction between their behaviors and objectives. In particular, the traditional spatial voting models rule out the influence of *party ideologies* upon citizens' preferences. Nevertheless, in countries where two parties dominate the political sphere, empirical evidence suggests that citizens' ideological preferences are often driven by their *party identifications*[2].

When party identification is strong it often implies voting for "one's" party's nominee (*straight ticket voting*). However, recent empirical evidence suggests that an interpretation of party identification that allows for different strengths and direc-

[2]See e.g. Bartels (2002), Evans and Andersen (2004), Goren (2005), Carsey and Laymain (2006), and Dancey and Goren (2010). This is particularly salient in the U.S. See Milazzo, Adams and Green (2012) for an account of its lower salience in British politics.

tions may be more appropriate to explain citizens' behavior.[3] Notably, citizens with moderate or mixed views seem to be of particular relevance. For instance, in the United States, one of the most prominent cases of a two-party system in which party identification is an important predictor of voting behavior, about half of the citizens declare themselves to be moderates or are unable to place themselves on an ideological scale (see e.g. Fiorina, Abrams and Pope 2006; Fiorina and Abrams 2008; Fiorina, Abrams and Pope 2008). According to the Pew Research Center (2014b) data on the responses to 10 ideological values questions, 39% of Americans take a roughly equal number of liberal and conservative positions.[4] Evidence from this survey and other sources (e.g. Hetherington 2008) also suggests that these citizens are less likely to turnout than strong partisans and that their voting behavior is less consistent than the one of citizens that have a strong identification to a party.[5]

In this chapter, we build a simple spatial voting model consistent with these observations. We consider two cases in terms of the relationship between candidates and party ideologies. In the first, *non-ideological candidates* maximize plurality independent of party ideologies. Candidates are only interested in the electoral outcome and can adopt any available strategy in the political domain. In the second case, each candidate cares about the ideology of "his" party.[6] In this case, *ideological candidates* maximize plurality bounded to the strategies close to the ideology of their respective party. The aim of the chapter is then twofold: (i) to put forward an extended spatial voting model, based on party ideologies and citizens' identifications, to study turnout and voting behaviors; and (ii) to characterize the electoral equilibrium of a two-candidate electoral competition in a political domain with two parties. By do-

[3] See Katz (1979) and Weisberg (1980) for early accounts of the "multidimensional" nature of party identification.

[4] This is down from about half of the American public in surveys conducted in 1994 and 2004. Meanwhile, the overall share of American citizens who express consistently liberal or consistently conservative opinions has doubled over the past two decades from 10% to 21%.

[5] In general, partisans are more likely to vote than "mixed voters", follow more regularly the news, are more interested in politics, and participate more than mixed voters in political activities (see Pew Research Center 2014b). For instance, while 78% and 58% of consistent conservatives and consistent liberals say they always vote respectively, only 39% of those who hold a mix of conservative and liberal views describe themselves as regular voters.

[6] To avoid awkward wording, we refer to candidates in the masculine and citizens in the feminine.

ing so, we provide a unified theory of turnout and voting behaviors with meaningful predictions about citizens' and candidates' behaviors.

We consider a multidimensional ideological space and represent party ideology as a (preferred) *ideological point* and a (surrounding) *acceptance region*. Any position within this region is perceived to be accepted by the party and any position outside (of it) is perceived to be rejected. The fixed and exogenous nature of party ideology intends to capture the parties' ideological "coherence and immobility" as defended by Downs (1957). Similarly, these ideologies can be thought to represent the "silent ideology" of commonly held interests shared by the *party's elite* (see Flinn and Wirt 1965). Finally, this representation is also consistent with the ideological views of other social groups that influence turnout and voting behaviors such as religion and ethnicity. The model is then suitable to study the effects of the ideologies *and* the identifications on diverse social identity groups on political behavior.

We follow Hershey (2015) and see party identification as a *social identity motive*. Taking parties or groups' ideologies as reference points, citizens form their spatial preferences based on the *ideological cues* (the ideological point and acceptance region) provided by their single *or* multiple identifications. That is, citizens can identify with, and focus upon the ideological cues of one *or* two parties simultaneously. On the one hand, we represent (unconflicted) *partisans* as citizens that identify with a single party. On the other hand, we represent *conflicted partisans* as citizens that identify with two parties. A citizen anticipates an *identity gain* in voting for a candidate that is accepted by a party she identifies with. Voting is anticipated to be costly if it implies the *betrayal* of any of the citizen's party identifications: if by the act of turnout and voting a citizen would vote for a candidate that adopts a position that is not accepted by *all* the parties she identifies with. In this sense, citizens are *betrayal averse*. They wish to be able to support a position that respects the single ideology or reconciles the multiple ones they believe in. These are interpreted as *intrinsic* gains and losses that can mobilize a citizen to turn out or abstain, independent of her probability to influence the outcome of the election.

To illustrate, take the example of the Democratic and Republican parties in the United States. These parties provide two reference points in terms of values and worldviews that influence citizens' stands on different issues. Citizens use their identifications to focus on either the liberal (Democratic) worldview, the conservative (Republican) worldview, or both. In this sense, Democrats focus on and adopt a liberal ideology while Republicans adopt a conservative one (what has been referred to by Levendusky 2009 and others as the *partisan sort*). Citizens with mixed or moderate views, on the other hand, hold seemingly conflicting views that are congruent with both ideologies. In our model this conflict is resolved in favor of a (*weak*) identification with the two ideologies: Conflicted partisans form the ideological preferences that give rise to their voting behavior based on the ideological cues given by the two parties. These citizens wish, if voting, to be able to reconcile both views and endorse a compromise between the two ideologies. If this is not possible, then voting is associated with a sense of betrayal and a psychological cost.

With two parties, this implies that an individual citizen turns out and votes if and only if there is a candidate position that is accepted by all the parties she identifies with. Then, if this is the case, a partisan votes for the candidate that is closest to the ideological point of the single party she identifies with. This is the available choice that procures her the highest identity gain.[7] For conflicted partisans, their ideological preferences depend upon the ideologies of the two parties. We show that a *conflicted voter's curse* emerges: If there is no position that reconciles the ideological views of both parties, it is always rational for conflicted voters to abstain even if they are, as a group, a majority. This curse is due to the high degree of perceived polarization in party ideologies, and holds no matter what position either of the candidates adopts.

[7]This means that under some circumstances, a partisan of one party may vote for the candidate affiliated to a different party. To substantiate this possibility, one may recall again the example of the United States, in which most, if not all elections, have featured large numbers of partisans voting against their party's nominee (Green, Palmquist and Schickler 2002). The so-called "Reagan Democrats" were an expression of that. In the U.K., "Essex Men" was the connotation given to the citizens that voted across their partisanship in the 1979, 1983, and 1987 general elections.

Using this approach, we address the questions of whether an electoral equilibrium exists, where such an equilibrium is located, and how the location is related to the ideological positions and acceptance regions of the two parties. We show that with non-ideological candidates, a small number of (pure strategy Nash) equilibrium pairs exist for most preference distributions. In particular, candidates converge to the position of the "median voter" if an *overlap region* between the parties' acceptance regions exists. If an overlap region does not exist, candidates converge as long as the partisans of one party are more numerous than the partisans of the other. This suggests that candidates behave independently of conflicted partisans in the absence of shared ideological views, even if they would otherwise be pivotal for the electoral outcome.

Ideological candidates converge if and only if an overlap region exists *and* conflicted partisans are pivotal. If conflicted partisans are not pivotal but an overlap region exists, the *majoritarian candidate* - the candidate affiliated to the party with which a majority of partisans identifies with - can adopt any position around the ideology of his party. The *minoritarian candidate* moves towards the center of the political domain to gain the support of conflicted partisans even though he is certain to lose the election. This means that majoritarian candidates enjoy more strategic flexibility than minoritarian candidates. Finally, if an overlap region does not exist, ideological candidates can adopt any position within the acceptance region of their respective party. This indicates that the absence of shared ideological views leads ideological candidates to adopt quite unpredictable behaviors in our setting.

Our model intends to illustrate the behavior of some of the voters that compose the political center, and how it could influence candidates' positions in an election where two groups or parties dominate the political sphere. In our setting, a high degree of perceived party polarization deters turnout as well as the convergence of candidates to the center of the political domain. In particular, turnout increases and candidates converge to a moderate position if and only if the parties share ideological views *and* conflicted partisans are pivotal. The results of this chapter pertain to be

illustrative and bring some insights into possible relationships between party ideologies, party identifications, polarization, citizens' and candidates' behavior. But care should be taken in the interpretation of these results, since our model relies upon demanding assumptions. For example, citizens have no policy preferences that are independent of their party identifications in our framework. Similarly, many voters that compose the "political center" may have policy preferences that reflect a mix of the views of the two parties (e.g. the economic view of the Republican party and the Democratic party's view on social issues). Our framework proposes a way to model party ideologies that could be used to represent such "eclectic voters", but our current version of the model illustrates the behavior of only one potential "type" of voters that may compose the political center that are not eclectic in this sense. We return to these issues when we discuss some of the limitations and extensions of our framework in Section 4.4.2.

One of the main purposes of the model is to explore, in the Downsian tradition, a rational intrinsic motivation that reconciles the spatial theory of voting with positive turnout rates. In this sense, it is most related to theories that have used *non-consequentialist motivations* to rationalize voting and turnout behaviors in a single framework.[8] For instance, Brennan and Hamlin (1998) analyze a spatial voting model in which turnout and voting behaviors are rationalized by the will to *express* support for one or the other candidate. Other authors have considered "relational goods" and the strategic calculus of groups as the major drivers of both turnout and voting behaviors (e.g. Morton 1987, 1991; Uhlaner 1989). Still others have proposed rationalizations based on party activism (e.g. Aldrich 1983, 1989). To the best of our knowledge, there is no previous spatial voting framework that uses (multiple) social

[8]These models, like ours, are in part motivated by the fact that the rational voting model is not consistent with positive turnout rates and that exogenous explanations of turnout (e.g. an individual *sense of civic duty* in Riker and Ordeshook 1968) deprive the rational model of a coherent and predictive rationality (Aldrich 1993; Green and Shapiro 1994). See Dhillon and Peralta (2002), Mueller (2003), Feddersen (2004), and Geys (2006) for general reviews. See Shayo and Harel (2012) for a recent discussion and experimental evidence on non-consequentialist voting.

identifications to explain voting *and* turnout behaviors.[9] Our theoretical framework also differs from previous contributions in the way that it models party ideologies and the focus on conflicted voters' behavior. Consequently, both the exogenous parameters and empirical predictions of our model remain significantly different from previous studies.

The remainder of the chapter is organized as follows. The next section is devoted to the formal model. In Section 4.3 we present the equilibrium results for the two-candidate electoral competition with two parties. We divide Section 4.3 into two subsections: In the first we characterize the electoral equilibrium with non-ideological candidates, and in the second we characterize it with candidates that care about ideology. We discuss the underlying behavioral theory as well as limitations and extensions to the model in Section 4.4. We summarize the implications of our analysis in Section 4.5.

4.2 Model

In this section we introduce our model of turnout and voting behaviors. In subsection 4.2.1 we introduce the general setting. Subsection 4.2.2 is devoted to party ideologies. In subsection 4.2.3 we present the spatial preferences of citizens based on their party identifications; and in subsection 4.2.4 we introduce citizens' turnout and voting decisions. The candidates' objectives are presented together with the analysis of the electoral equilibrium in Section 4.3.

4.2.1 Setting

Let $X \subseteq \mathbf{R}^m$ denote a set of alternatives such that each $x \in X$ is a column vector $x = (x_1, ..., x_m)$. We interpret these alternatives as vectors of positions on *policy issues*, such as the level of income tax, as well as on *non-policy issues* (under the control of

[9] See Shayo (2009) for a model that uses social identity to rationalize the political economy of income redistribution.

candidates), such as perceived image or personality.[10] This excludes non-policy issues that are not under the control of candidates such as age or ethnicity. We call X the **political domain**.

We consider a finite set of N **citizens** and an electoral competition between **two candidates** 1 and 2. Let $\theta_1 = (\theta_{11}, ..., \theta_{1m})$ and $\theta_2 = (\theta_{21}, ..., \theta_{2m})$ denote the **strategies** that candidates choose in the political domain. As in the traditional spatial models of electoral competition, we assume that (1) all citizens have identical perceptions of candidates' strategies and that (2) candidates know citizens' preferences.

4.2.2 Party Ideologies

There are two **parties** b and r, the **blue party** and the **red party** respectively. A party $p \in \{b, r\}$ is characterized by two parameters. The first of these is an **ideological point** δ_p in the political domain:

P 1. *For all* $p \in \{b, r\}$, $\delta_p = (\delta_{p1}, \delta_{p2}, ..., \delta_{pm}) \in X$.

The ideological point δ_p represents the citizens' (and candidates') estimation of the party *p*'s ideological preferred position on each of the m dimensions. This shared perception can be seen as the reference ideological point socially attached to the party. In this sense, the ideological point can be thought of as the public view of the party's coherent set of stands on different issues. Alternatively, it can be interpreted as the "silent ideology" of commonly held interests shared by the "like-minded men and women" that run the party (see Flinn and Wirt 1965). The second parameter is a (political) **acceptance region** A_p delimited by a distance $d_p \in R^{++}$ to δ_p:

P 2. *For all* $p \in \{b, r\}$, $A_p(\delta_p, d_p) = \{x \in X : ||x - \delta_p|| \leq d_p\}$.

where $||\cdot||$ denotes the Euclidean distance. For each $p \in \{b, r\}$, the acceptance region A_p represents the positions within a "threshold" distance d_p that are perceived to be accepted by the party or acceptable with respect to the party's ideology. The

[10] See Hinich and Ordeshook (1969) and Brennan and Hamlin (1998) for discussions about the interpretation of the various dimensions on models of spatial voting.

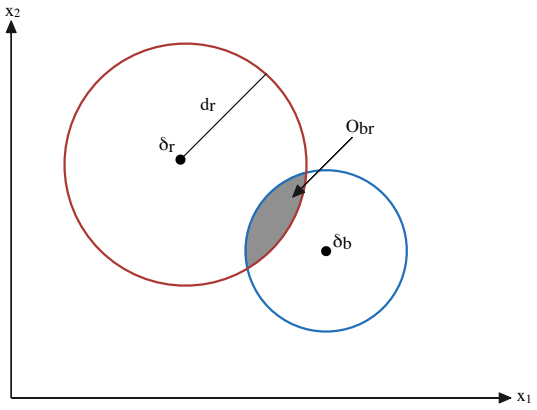


Figure 4.1 – Party ideologies with a non-empty overlap region.

different threshold distances can be interpreted, for instance, as the perceived "range of opinion" allowed by the two parties. In this sense, a smaller acceptance region corresponds to a party being perceived as more demanding, i.e., a party that is perceived to accept less discrepancy with respect to its preferred ideological position.

This means that whenever a candidate adopts a position within this distance it is perceived as acceptable with respect to the party's ideology, while if a candidate adopts a position outside this distance the party is perceived to reject the candidate's ideological stand. Then, it may be the case that a red ideological candidate adopts an ideological position that is in the acceptance region of the blue party. This is not to be interpreted as if the blue party supports the red candidate. Rather, it is interpreted *as if* the position adopted by the red candidate is perceived as an acceptable position according to the standards of the ideology of the blue party.

Finally, if there is an intersection between A_b and A_r we call it an **overlap region** and denote it by O_{br}. This region, if it exists, corresponds to the ideological positions in the political domain that are accepted by both parties. It can be interpreted as the ideological common ground or shared views of both parties: the positions which reconcile both ideologies. Figure 5.1 represents these concepts in a two-dimensional domain.

4.2.3 Citizens' Preferences

Each citizen $i \in N$ has complete and transitive preferences over X, that are represented by a **utility function** $u_i : X \to \mathbf{R}$. Let $P_i \subseteq \{b, r\}$ denote the subset of parties that citizen i identifies with. Then, for all $i \in N$, either $P_i = \{b\}$ (citizen i identifies with the blue party), $P_i = \{r\}$ (citizen i identifies with the red party), or $P_i = \{b, r\}$ (citizen i identifies with the two parties). We call **partisan** a citizen that identifies with a single party: from these, we distinguish **blue partisans** and **red partisans**. A citizen that identifies with the two parties is referred to as **conflicted partisan** (or **bi-partisan**).

The strength of these identifications is individual specific. It is numerically measured by the **weight(s)** $I_i^p \in]0, 1]$ that citizen i attaches to each party p she identifies with. The higher the I_i^p the higher the identification is interpreted to be.

A citizen evaluates alternatives according to her weighted party identifications. In order to specify these preferences, we proceed as follows. First, we distinguish the three *intrinsic* sources of utility benefits and losses with which a citizen evaluates the alternatives in the political domain.[11] Second, we combine these three sources to specify the utility of voting for a given position in X. The first source, $u_{ip}^A(x)$, is citizen i's (simple) **Euclidean preferences** with respect to the ideological points δ_p of *all* the parties she identifies with:

U 1. *For all $i \in N$, all $p \in P_i$, and all $x \in X$,*

$$u_{ip}^A(x) = -\frac{I_i^p}{d_p} \cdot ||\delta_p - x||.$$

U1 entails that positions closer to the preferred ideological point of "one's" party are preferred to positions which are farther from this point in what respects the identification with this party. This can be interpreted as the (psychological) loss associated with positions that do not fully represent one's identifications. On the

[11] They are intrinsic since citizens have an effective choice on if and how they vote, but they do not have an effective choice between alternative policy outcomes since their single vote has a negligible probability of being pivotal (see Brennan and Hamlin 1998, 155-6).

one hand, for a partisan this is a close formulation to the one used in traditional analysis of spatial voting. The main difference is that here the Euclidean preferences are not with respect to an "ideal point". Instead, they are with respect to her party's preferred ideological point. In this sense, the ideology of her party is used as a cue on the formation of her preferences. On the other hand, a conflicted partisan exhibits Euclidean preferences with respect to the ideological points of the two parties. The two parties provide ideological cues that a conflicted voter i weights according to (i) the individual weight I_i^p of each party and (ii) the size of the parties' acceptance regions given by d_p. First, all else being equal, any "move" towards the ideological point of a party of *higher weight* (i.e., of stronger identification) than another entails a higher utility gain than a similar move towards the ideological point of the party with *lower weight*. Second, all else being equal, any move towards the ideological point of a party with a *smaller acceptance region* (i.e., a more demanding party) than another entails a higher utility gain than a similar move towards the ideological point of the party with a *larger acceptance region*.

The remaining two sources correspond to the psychological gains and losses associated with following or betraying one's identifications. Voting is anticipated to generate identity gains or losses depending on whether or not the act of voting can be done in accordance with the perceived *behavioral prescriptions* of the parties/groups one identifies with. These prescriptions are given by the parties' cues (the ideological points and acceptance regions) and generate utility benefits and losses, $u_{ip}^B(x)$ and $u_i^C(x)$, as follows:

U 2. *For all* $i \in N$, *all* $p \in P_i$, *and all* $x \in X$,

$$u_{ip}^B(x) = \begin{cases} I_i^p \ \textit{if } x \in A_p \\ 0 \ \textit{otherwise.} \end{cases}$$

U 3. *For all* $i \in N$, *all* $p \in P_i$, *and all* $x \in X$,

$$u_i^C(x) = \begin{cases} -c_i \ \textit{if} \ x \notin \cap_{p \in P_i} A_p \\ 0 \ \textit{otherwise.} \end{cases}$$

with $c_i > I_i^p$ for all $p \in P_i$. U2 can be interpreted as the **identity gain** that a citizen receives if she is to vote for a candidate that adopts a position that is accepted by a party she identifies with. The magnitude of this benefit depends upon the weight the party has for her. Note that this gain can be associated with following the prescription of a single party in the case of partisans, or of one *or* two parties in the case of bi-partisans. U3 represents the **cost of betrayal** that a citizen may incur with the act of voting. In particular, voting is anticipated to be costly if it implies the *betrayal* of any of the citizen's party identifications: if by the act of turnout and voting a citizen would vote for a candidate that adopts a position that is not accepted by *all* the parties she identifies with.

We assume the cost of betrayal to be greater than the identity gain associated with voting for a candidate that adopts a position that is accepted by a given party (reflected in $c_i > I_i^p$ for all $p \in P_i$). This means that citizens are *betrayal averse*: they give higher weight to identity losses than to identity gains. This is in accordance with the strong evidence on loss aversion on pecuniary/monetary items (e.g. Tversky and Kahneman 1991; Bowman, Minehart and Rabin 1999) and with the mild evidence in support of its extension to non-monetary effects (e.g. Galanter and Pliner 1974; Crockett, Kurth-Nelson, Siegel, Dayan and Dolan 2014).[12]

For conflicted partisans, this entails that they anticipate to bear a psychological cost whenever they cannot vote without betraying *one* party but not necessarily the other. Then, in the case of abstention there may be a foregone identity gain for not having voted for a candidate that is accepted by one of the parties a citizen identifies with. One way of making sense of this is that while betrayal is associated

[12]For instance, Crockett et al. (2014) show that people require more compensation to increase pain (in the form of small shocks) than they are willing to pay to decrease it by the same amount. See Dhar and Wertenbroch (2000) and Vendrik and Woltjer (2007) for loss aversion for hedonic and utilitarian attributes and on life satisfaction with respect to relative income respectively.

with anticipated regret of an *action* against one's identification, the foregone identity gain is associated with anticipated regret of *inaction* (not having voted for the given candidate). For that reason, betrayal aversion is in accordance with the evidence that the regret experienced from action is more intense than that from inaction or omission (e.g. Kahneman and Tversky 1982; Crockett et al. 2014). Additionally, this assumption captures the intuition that these citizens want to support an outcome that can be reconciled with both ideologies they have a personal attachment to. In other words, conflicted partisans find ideological acquaintances in the worldview of the two parties and wish, if voting, to endorse a compromise between the two. This seems a sensible assumption to represent the preferences of voters with mixed and moderate views that have a weak and not very disparate identification to both parties. It would be less adequate to illustrate the preferences of conflicted voters that attach very different weights to the two parties.[13]

For all $p \in P_i$, citizen i evaluates all alternatives $x \in X$ according to the combination of these three intrinsic utility sources, such that $u_i(x) = \sum_{p \in P_i}[u^A_{ip}(x) + u^B_{ip}(x)] + u^C_i(x)$. Then, citizen *i*'s utility of voting for (a candidate at) position x can be written as follows:

$$\begin{aligned} u_i(x) &= \sum_{p \in P_i} (I^p_i - \frac{I^p_i}{d_p} \cdot ||\delta_p - x||) \text{ if } ||\delta_p - x|| \leq d_p \text{ for all } p \in P_i \\ &= \sum_{p \in P_i} (-\frac{I^p_i}{d_p} \cdot ||\delta_p - x||) + I^q_i - c_i \text{ if } ||\delta_q - x|| \leq d_q \text{ for only one } q \in P_i \text{ and } \#P_i = 2 \\ &= \sum_{p \in P_i} (-\frac{I^p_i}{d_p} \cdot ||\delta_p - x||) - c_i \; otherwise. \end{aligned} \tag{4.1}$$

[13] A weaker version of betrayal aversion would be one that would hold for each party separately (i.e., $c^p_i > I^p_i$ for all $p \in P_i$). This is equivalent with respect to partisans. However, it would change the behavior of bi-partisans. Below we comment on the differences that would follow from this weaker aversion.

4.2.4 Turnout and Voting Decisions

For any combination of strategies (θ_1, θ_2) chosen by the two candidates, citizen i either votes for one of these two candidates or abstains. Let $t_i \in \{0, 1\}$ and $v_i \in \{1, 2\}$ represent citizen i's turnout and voting decisions respectively. The **turnout decision** t_i takes the value $t_i = 1$ if citizen i participates in the election and $t_i = 0$ otherwise. The **voting decision** v_i available to citizen i in case of turnout is between candidate 1 ($v_i = 1$) and candidate 2 ($v_i = 2$). Then, citizen i solves the following maximization problem:

$$\max_{t_i \in \{0,1\},\ v_i \in \{1,2\}} t_i \cdot [u_i(\theta_{v_i})] \tag{4.2}$$

This decision problem can be seen as a two-stage optimization problem, where in the first stage the citizen decides whether or not to participate in the election and in the second stage she decides whom to vote for. Solving it backwards, it is straightforward to see that citizen *i*'s optimal voting decision is to vote for the candidate $k \in \{1, 2\}$ that proposes the alternative associated with higher utility:

$$v_i(u_i, \theta_k) = \begin{cases} 1 & \text{if } u_i(\theta_1) > u_i(\theta_2) \\ 2 & \text{if } u_i(\theta_1) < u_i(\theta_2). \end{cases} \tag{4.3}$$

and in case of indifference to vote with equal probability for either candidate. Indeed, if a citizen is indifferent between two candidates that are accepted by the party/parties she identifies with, there is no reason for this citizen to abstain. Instead, she can maximize her utility by randomly selecting one of the two candidates: much as the hot and hungry sun-bather who is close and equidistant from two ice-cream sellers chooses randomly from whom to buy an ice-cream (see Brennan and Hamlin 1998, 157). Then, it follows from (4.1), (4.2), and (4.3) that citizen *i*'s optimal turnout decision is to decide whether or not to participate in the election according to the following rule:

Proposition 1. *For all* $i \in N$ *and* $k \in \{1, 2\}$*:*

$$t_i(u_i, \theta_k) = \begin{cases} 1 & \text{if } \exists k \text{ such that } ||\delta_p - \theta_k|| \leq d_p \text{ for all } p \in P_i \\ 0 & \text{otherwise.} \end{cases}$$

Proof. See appendix (for all proofs). □

In words, a citizen turns out if and only if there exists a candidate that is accepted by *all* the parties she identifies with.[14] Otherwise, the citizen abstains. This implies that the possible candidates' positions for which a partisan may cast a vote for are within the acceptance region of her party. In the case of a bi-partisan, the possible candidates' positions for which she may cast a vote for are within the intersection of the acceptance regions of the two parties, i.e., within the overlap region O_{br}. This underlies the conflicted voter's curse: If there is no position that reconciles the ideological views of both parties, it is always rational for conflicted partisans to abstain. This behavior is rational irrespective of candidates' positions, and even if conflicted partisans are, as a group, a majority. This means that conflicted partisans may abstain even though they could, together, change the outcome of the election for their preferred outcome.[15]

The conflicted voter's curse (and partisans' turnout behavior) are direct implications of betrayal aversion. Betrayal aversion is a different reason for abstention than the most common *indifference* and *alienation* hypotheses.[16] In particular, a citizen abstains if all the candidates are sufficiently far from the *ideological points* of *all* the

[14] We have implicitly assumed that in the case that a citizen is indifferent between turning out and abstaining the tie is broken in favor of participation; this entails no loss of generality.

[15] If one assumed instead $c_i^p > I_i^p$ for all $p \in P_i$ (see footnote 12), then the conflicted voter's curse would hold if $c_i^b \geq I_i^r$ and $c_i^r \geq I_i^b$ for all conflicted partisans. Otherwise, it would hold or not depending upon the adjusted weights that each conflicted partisan attached to each party (as would be the case if one assumed $c_i^p = I_i^p$ for all $p \in P_i$ or $c_i^p < I_i^p$ for all $p \in P_i$). Accordingly, these alternative specifications would entail less general analytic results concerning both the citizens' and candidates' behaviors. Nonetheless, to ascertain the contextual determinants responsible for the existence and form(s) of betrayal aversion is a relevant and open empirical question.

[16] Under indifference a citizen abstains if all the candidates assume a sufficiently similar position and there is not sufficient difference in the payoffs of voting for each candidate (see e.g. Hinich, Ledyard and Ordeshook 1972). Under alienation a citizen abstains if all the candidates are sufficiently far from her ideal point (see e.g. Hinich and Ordeshook 1969).

parties she identifies with. Then, partisans abstain much as citizens abstain in accordance to alienation (though their Euclidean preferences are with respect to their party's preferred ideological point), while conflicted partisans feel alienated as soon as they cannot, through voting, resolve their conflict due to multiple identifications. If there is no position available that represents a compromise and reconciles the ideologies of the two parties, then all positions seem unsatisfactory and conflicted partisans abstain.

4.3 Electoral Equilibrium

In this section we look at the consequences of party ideologies and citizens' preferences on the choices that candidates offer to the citizens. In subsection 4.3.1 we characterize the candidates' equilibrium strategies when candidates are non-ideological, i.e., when they are unrestricted in their strategy choices. In subsection 4.3.2 we characterize the candidates' equilibrium strategies when candidates are ideological, which is reflected in restricted choices on the basis of their party affiliations.

Let $N_b = \#\{i : \ P_i = \{b\}\}$, $N_r = \#\{i : \ P_i = \{r\}\}$, and $N_{br} = \#\{i : \ P_i = \{b, r\}\}$. Note that $N_b + N_r + N_{br} = N$. We impose the two following conditions in our analysis:

A 1. *$N_b > 0$, $N_r > 0$ and $N_{br} > 0$.*

A 2. *For any $i \in N_{br}$, $\frac{I_i^b}{d_b} \neq \frac{I_i^r}{d_r}$.*

The first condition requires that there exists at least one blue partisan, one red partisan, and one bi-partisan. The second condition requires that the weight of the two parties, when adjusted by the size of the acceptance region, be different for all conflicted partisans. It captures the intuitive idea that it is possible to distinguish bi-partisans according to the party they lean towards (e.g. citizens who lean Democratic and citizens who lean Republican). None of these conditions change our results substantially, but they entail considerable gains in parsimony.

We start our analysis by establishing the circumstances under which a citizen i has a "satisfactory" ideological preferred point x_i^* (*bliss point*), and, if it exists, where that bliss point (the point that gives the highest utility) is located. Such a point is satisfactory in the sense that voting for a candidate in that position entails a greater utility than abstaining. Unconflicted and conflicted partisans differ in this respect. Let $\alpha_b, \alpha_r \in O_{br}$ denote the points within any non-empty overlap region that minimize the distance to δ_b and δ_r respectively. Then:

Lemma 1. *(a.1) For all $i \in N_b \cup N_r$ (partisans) there always exists $x_i^* \in X$ s.t. $x_i^* = \underset{t_i \in \{0,1\},\ x \in X}{argmax}\ t_i[u_i(x)]$;*

(a.2) For all $i \in N_b$ (blue partisans) $x_i^* = \delta_b$ and for all $i \in N_r$ (red partisans) $x_i^* = \delta_r$;

(b.1) For all $i \in N_{br}$ (bi-partisans) there exists $x_i^* \in X$ s.t. $x_i^* = \underset{t_i \in \{0,1\},\ x \in X}{argmax}\ t_i[u_i(x)]$ if and only if $O_{br} \neq \emptyset$;

(b.2) If $O_{br} \neq \emptyset$, $x_i^* = \alpha_b$ for all $i \in N_{br}$ s.t. $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$ and $x_i^* = \alpha_r$ for all $i \in N_{br}$ s.t. $\frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}$.

Lemma 1 shows several interesting features of citizens' behavior implied by our assumptions. First, partisans always have the same bliss point irrespective of the parties' polarization. This point is at the preferred ideological position of their party. Second, conflicted partisans only have a satisfactory ideological preferred position if the overlap region is not empty. When this is the case, Lemma 1 shows that the bliss points of conflicted partisans are in between the preferred ideological positions of the two parties. This means that whenever party polarization is weak, all citizens have a satisfactory ideological preferred position that takes one of four points in the political domain. Figure 4.2 illustrates the positions of citizens' bliss points in a two-dimensional political domain.

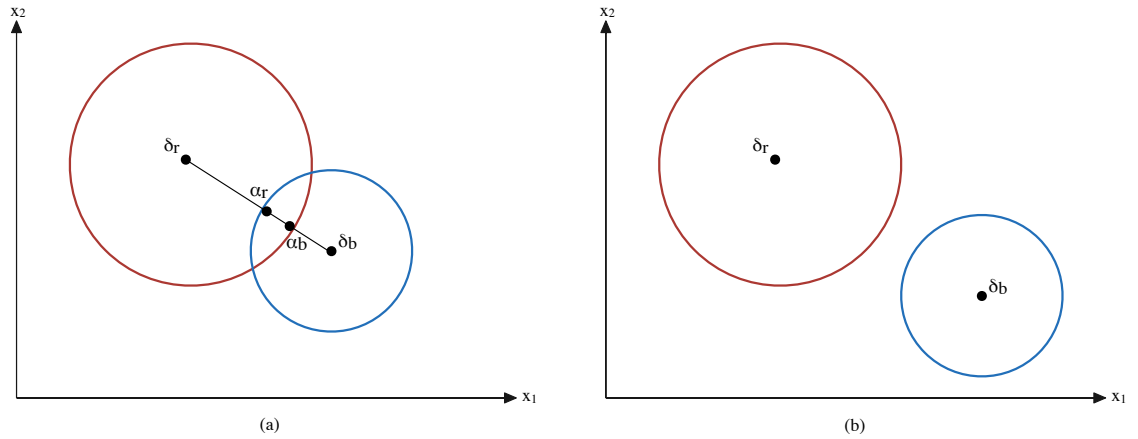


Figure 4.2 – (a) Bliss points with an overlap region; (b) Bliss points with no overlap region.

4.3.1 Non-ideological Candidates

With unrestricted strategies, candidates can adopt any position in the political domain. In our framework, this means that candidates have "no color". This is the standard assumption in spatial models of electoral competition. Let $V_k(\theta_1, \theta_2)$ denote the number of citizens that vote for candidate k and $Pl_k(\theta_1, \theta_2)$ denote candidate k's (expected) **plurality** given candidates' strategies θ_1 and θ_2. Then, $Pl_1(\theta_1, \theta_2) = V_1(\theta_1, \theta_2) - V_2(\theta_1, \theta_2)$ and $Pl_2(\theta_1, \theta_2) = V_2(\theta_1, \theta_2) - V_1(\theta_1, \theta_2)$. Candidates choose a strategy from X that maximizes plurality.[17] We consider that the candidates' payoffs are their (expected) pluralities, which yields a zero-sum game since $Pl_1(\theta_1, \theta_2) = -Pl_2(\theta_1, \theta_2)$. Then, a strategy combination (θ_1^*, θ_2^*) is a (pure strategy Nash) **equilibrium pair** if and only if:

$$Pl_1(\theta_1^*, \theta_2) \geq Pl_1(\theta_1^*, \theta_2^*) \geq Pl_1(\theta_1, \theta_2^*) \text{ for all } \theta_1, \theta_2 \in X$$

and if $(\theta_1^*, \theta_2^*) = (\theta^*, \theta^*)$ then (θ^*, θ^*) is said to be a **convergent equilibrium**. When the number of citizens is odd, the electoral competition game admits a convergent equilibrium as follows:

[17] This is consistent with a plurality rule election in which each citizen is allowed to vote for one and only one candidate and the candidate with most votes wins the election. See Aranson, Hinich and Ordeshook (1974) for a discussion of different candidates' objective functions and support for the rationality of maximizing expected plurality in a two-candidate plurality rule election.

Theorem 1. *If N is odd and $O_{br} \neq \emptyset$, then there exists a convergent equilibrium (θ^*, θ^*) in one of the strategies $\theta^* \in \{\delta_b, \delta_r, \alpha_b, \alpha_r\}$.*

This means that when an overlap region exists and N is odd then candidates converge to one of the four potential positions where the citizens' bliss points may be located. Which one depends upon the position of the "median voter": That point is a Condorcet winner from the candidates' point of view, i.e., a point that wins or ties against any other alternative with majority rule. As it can be seen from Figure 4.2.a., this result hinges on the fact that all $x^* = \alpha_b$, α_r, δ_b, δ_r are on a line segment. The location of the convergent equilibrium is restricted to one of the four positions depicted in the figure.

There are cases for which an overlap region exists and a convergent equilibrium is not the solution of the electoral competition game. Still, it is possible to identify circumstances under which a limited number of equilibrium pairs exist. The essential condition is that there is a majority made of partisans of one party and bi-partisans that lean towards this party (e.g. strong and lean Democrats are a majority):

B 1. $N_b + \#\{i \in N_{br} : \frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}\} \neq N_r + \#\{i \in N_{br} : \frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}\}$.

This condition allows us to show that whenever an overlap region exists and candidates are unrestricted in their strategy choices then the existence of a small number of equilibrium pairs is quite general:

Theorem 2. *Suppose $O_{br} \neq \emptyset$. Then, there exists at least one and no more than four equilibrium pairs (θ_1^*, θ_2^*) such that $\theta_1^*, \theta_2^* \in \{\delta_b, \delta_r, \alpha_b, \alpha_r\}$ if and only if B1 holds.*

Condition B1 is necessary and sufficient for Theorem 2. Nevertheless, if B1 does not hold, the set of dominant positions is still considerably restricted. Although it is not possible to determine a precise number of equilibrium pairs, the location of the equilibrium pairs is restricted to the points within the overlap region that are on the line segment that connects the ideological points of both parties δ_b and δ_r. All these points can guarantee a tie against each other (for some cases one of the ideological points as well), and candidates can adopt any of these positions at equilibrium.

If instead there is no overlap region, candidates almost always converge even though it is not to the position of the "median voter". In this case, the electoral competition game admits a convergent equilibrium as follows:

Theorem 3. *Suppose $O_{br} = \emptyset$. Then, there exists a convergent equilibrium (θ^*, θ^*) in one of the strategies $\theta^* \in \{\delta_b, \delta_r\}$ if and only if $N_b \neq N_r$.*

In words, if there is no overlap between the two ideologies then candidates converge as long as the number of blue partisans is *not equal* to the number of red partisans. Their point of convergence is the preferred ideological point of the party that a greater number of citizens identify with (see Figure 4.2.b). Otherwise (in the unlikely case that $N_b = N_r$), candidates may converge to the preferred ideological point of one party or diverge by each adopting a distinct position at one of the two parties' ideological points δ_b and δ_r.

An empty overlap region can be interpreted as a case of high degree of polarization among the perceived ideologies of the two major parties. In particular, it corresponds to a case in which no position reconciles the ideological views of both parties. This result suggests that if candidates are not restricted on their strategy choices, then high degrees of polarization may not stop them from converging to the preferred ideological point of one of the parties in order to maximize plurality. Turnout decreases not only due to the conflicted voters' abstention, but also due to the abstention of the partisans that are in minority.

When candidates do not care about ideology, and contrary to the traditional Downsian models, our framework does not predict the existence of a unique equilibrium. In this sense, our model rationalizes the empirical observation that candidates may diverge even if they are only interested in maximizing plurality. At the same time, our model captures the lower participation of citizens with moderate preferences that in the traditional spatial voting models with abstention would be the most likely to vote. And contrary to exogenous explanations such as a sense of civic duty, the social identity motive in our model rationalizes different levels of turnout en-

dogenous to candidates' positions. Before discussing the limitations and potential extensions of this framework, we turn to the case of candidates that have themselves ideological preferences.

4.3.2 Ideological Candidates

In the previous subsection, we have kept the assumption of most spatial voting models in terms of the absence of restrictions upon candidates' strategy choices. However, the assumption that candidates can adopt any position in the political domain can be unsatisfactory in many circumstances (see e.g. Downs 1957; Kramer and Klevorick 1974; Matthews 1979; Samuelson 1984).[18] Among other examples, candidates' party affiliations/identifications is one that seems rather salient. In many countries candidates are formally linked to parties and depend upon their support. Candidates may also want to be, and gain from being, ideologically coherent, credible, or loyal to *their* party. This is the case in the United States; it is also in many European countries, as the evidence from the Manifesto Research Group suggests: according to their comparative coding of party platforms in 19 European countries, candidates often vary their policies within ideologically delimited areas.[19]

Our framework conveys a natural way to restrict the set of possible strategies available for candidates that is consistent with these observations: each candidate is affiliated with one party and his strategies are restricted to the acceptance region of this party. Formally, we let candidates 1 and 2's **strategy spaces** $X_1, X_2 \subseteq X$ be $X_1 = A_b$ and $X_2 = A_r$. This can be interpreted *as if* candidate 1 is affiliated with party b and candidate 2 is affiliated with party r. We refer to them as the **blue** and **red candidate** respectively.

[18] For instance, in order to maintain credibility candidates may be constrained to strategies near a previously adopted position (Samuelson 1984), or restricted to strategies near a *status quo* such as a convergent equilibrium (Kramer and Klevorick 1974 and Coughlin and Nitzan 1981).

[19] See e.g. Budge, Robertson and Hearl (1987) for one of the original studies and Adams (2001) for a review.

Like non-ideological candidates, ideological candidates are interested in maximizing plurality. But now they choose a strategy to maximize this objective from X_k that is different from X. That is, they maximize plurality bounded by "their color". This means that a candidate can present himself to the election to maximize his plurality, even though, *a priori*, he has no possibility of winning the election. This seems a rational behavior in line with the conflict between maximizing plurality and ideological coherence. We consider that the (expected) pluralities are the candidates' payoffs, which yields again a zero-sum game. Now, a strategy combination (θ_1^*, θ_2^*) is a (pure strategy) equilibrium pair if and only if:

$$Pl_1(\theta_1^*, \theta_2) \geq Pl_1(\theta_1^*, \theta_2^*) \geq Pl_1(\theta_1, \theta_2^*) \text{ for all } \theta_1 \in X_1 \text{ and all } \theta_2 \in X_2$$

and if $(\theta_1^*, \theta_2^*) = (\theta^*, \theta^*)$ then (θ^*, θ^*) is also said to be a convergent equilibrium. The next result requires that we specify two additional conditions. First, conflicted partisans need to be **pivotal**; this means that they are numerous enough, as a group, to modify the result of the election:

C 1. $N_{br} \geq |N_b - N_r|$.

Second, it is not the case that either of the parties' ideological points is within the overlap region. This excludes cases of very strong overlap between the two ideologies (as in Figure 4.3.a):

C 2. $\delta_b, \delta_r \notin O_{br}$.

When the number of citizens is odd, the electoral competition game between ideological candidates admits a convergent equilibrium under the following conditions:

Theorem 4. *Suppose N is odd and C2 holds. Then, there exists a convergent equilibrium (θ^*, θ^*) in one of the strategies $\theta^* \in \{\alpha_b, \alpha_r\}$ if and only if $O_{br} \neq \emptyset$ and C1 holds.*

In words, if the overlap is not very strong, candidates converge if and only if the parties share ideological views and conflicted partisans are pivotal. In this case, candidates adopt a moderate position at the center of the political domain that leads

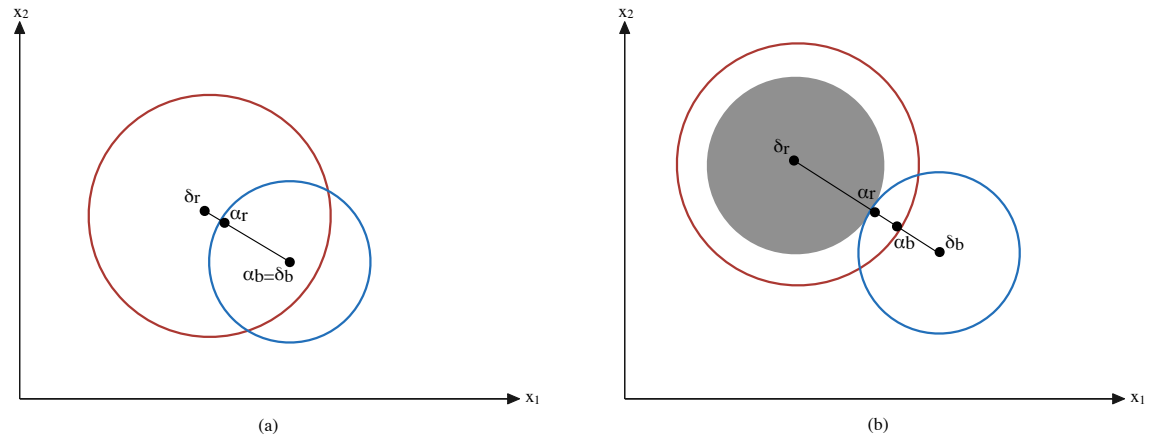


Figure 4.3 – (a) Strong overlap; (b) Weak overlap.

every citizen, including conflicted partisans, to turn out. Contrary to non-ideological candidates, the condition that conflicted partisans are pivotal is now essential for convergence. Indeed, Theorem 4's conditions are sufficient, and with an exception, necessary for convergence. The exception is a political domain characterized by a strong overlap between the two ideologies as in Figure 4.3.a. In this case, a convergent equilibrium may still exist even if conflicted partisans are not pivotal. In the figure's example, this would be the case if the blue partisans were a majority. In that case, both candidates would converge to the blue party's ideological point δ_b. For any other case a convergent equilibrium does not exist. This means that, as expected, convergence is harder with ideological candidates than with non-ideological candidates.

Nonetheless, even if conflicted partisans are not pivotal it is still possible to determine a restricted set of equilibrium pairs. The next result requires that there exists at least one bi-partisan leaning towards one party and one leaning towards the other party:

C 3. $\exists i, j \in N_{br}$ *such that* $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$ *and* $\frac{I_j^b}{d_b} < \frac{I_j^r}{d_r}$.

Then, in the case of a weak overlap between the two ideologies ($O_{br} \neq \emptyset$ and C2 holds), the dominant positions are restricted but one of the candidates has considerable flexibility in his strategy choices:

Theorem 5. *Suppose* $O_{br} \neq \emptyset$, *C1 does not hold, C2 and C3 hold. Then, if* $N_b > N_r + N_{br}$ *the equilibrium pairs* (θ_1^*, θ_2^*) *are such that* $\theta_1^* \in ||\delta_b - \theta_1^*|| < ||\delta_b - \alpha_b||$ *and* $\theta_2^* = \alpha_b$. *If* $N_r > N_b + N_{br}$ *the equilibrium pairs* (θ_1^*, θ_2^*) *are such that* $\theta_1^* = \alpha_r$ *and* $\theta_2^* \in ||\delta_r - \theta_2^*|| < ||\delta_r - \alpha_r||$.

Since bi-partisans are not pivotal, then partisans of one party are the majority. Suppose that these are the red partisans. In this case, the blue candidate adopts strategy α_r in order to maximize his plurality, even though it will be negative. It is rational for this candidate to do so since he cannot move beyond the range of opinion perceived to be allowed by his party, and such a position at the center of the political domain guarantees the support of blue partisans and conflicted partisans against the dominant positions of the red candidate. Indeed, the red candidate can adopt any position within A_r as long as this position is not farther from δ_r than α_r. Any of these strategies guarantees the support of the red partisans that is sufficient to win the election. These are the positions within the grey region in Figure 4.3.b.

This illustrates an interesting feature of the electoral competition between two ideological candidates when conflicted voters are not pivotal. While it is in the best interest of a minoritarian candidate to adopt a position at the center to gain the support of conflicted partisans, a majoritarian candidate has the flexibility to adopt several positions around the ideology of his party. On the one hand, the minoritarian candidate attracts *all* bi-partisans including the ones leaning towards the other party (e.g. voters who lean Republican vote for the Democratic candidate). On the other hand, the majoritarian candidate enjoys a majoritarian partisan core of support. This allows him to adopt, inclusively, what in our framework can be seen as more extreme positions (the points on the grey region North-West of δ_p in the example of Figure 4.3.b). This illustrates the potential for progressive behavior of majoritarian candidates and the more constrained and moderate behavior that minoritarian candidates may be induced to adopt.

If instead the parties are perceived to share no ideological views, then ideological candidates can adopt any position within the acceptance regions of their respective parties, i.e.:

Theorem 6. *Suppose* $O_{br} = \emptyset$. *Then, the equilibrium pairs* (θ_1^*, θ_2^*) *are such that* $\theta_1^* \in A_b$ *and* $\theta_2^* \in A_r$.

This means that in our setting candidates' behavior is quite unpredictable in cases in which no position is perceived to reconcile the views of both parties. It implies a causal arrow from strong party polarization to higher political/strategic flexibility of candidates. In particular, candidates may be able to adopt any position, even more extreme ones, as long as it is not perceived as firmly rejected by their party. This illustrates a potential relationship between party polarization, ideological coherence, and the possibility for candidates to take progressive or more extreme stands.

With ideological candidates, all citizens turn out when polarization is perceived to be weak (Theorems 4 and 5). Otherwise, turnout is composed of the partisans of each party (Theorem 6). Conflicted voters, due to the high degree of party polarization, abstain irrespective of the candidates' positions. It thus follows that conflicted partisans may even be a majority and candidates (ideological and non-ideological alike) may still adopt positions far away from the center of the political domain. Indeed, our results illustrate the possibility that when the perceived party polarization is strong candidates will favor positions closer to the preferred position of "their" partisans. This seems consistent with the cases in which candidates focus on "base voters", i.e., cases in which candidates ignore the political center in favor of their electoral core of support.

Our model rationalizes both centripetal and centrifugal forces within a limited set of possible positions. We modify several of the assumptions found in the traditional Downsian model. In this respect, we follow the several attempts that have been made to modify this model to make it compatible, among other things, with the evidence that candidates in most democratic countries generally adopt different po-

litical stands (e.g. Adams 2001). According to Grofman (2004, 40-1), who reviews and defends this so-called *neo-Downsian agenda*, it allows us to build towards an "institution-specific and voter preference-distribution-specific theory of party competition" that has testable implications in terms of comparative statics.[20] Since turnout and candidates' ideological positions vary from one election to the other, it is important to identify variables that differentiate elections. Our exercise is useful in this sense since it identifies the perception of the parties' ideological positions and their polarization, the distribution of citizens' identifications, and candidates' party affiliations as such variables. Different values of these parameters imply different testable implications concerning voting and turnout behaviors and candidates' centripetal or centrifugal movements.

Nevertheless, some caution should be exercised in the interpretation of these empirical implications. In the case of our model, they represent illustrative benchmarks of potential relationships between party ideologies, party identifications, polarization, citizens' and candidates' behavior. Indeed, we see our behavioral model mostly as a tool for gathering insights into these topics and our spatial framework as a tool for the development of new perspectives on how to model party ideologies and identifications. It is certainly not a full-fledged model, but we hope a step towards a more encompassing framework that takes these variables into account. It is also an attempt to model one type of voters that may compose the political center, suggesting future lines of research in this respect.

4.4 Discussion

We have advanced a *self-regarding motivation* to political behavior - a social-identity motive with individual identity gains and losses - that is consistent with *other-regarding*

[20] See Enelow and Hinich (1984) for an introduction and a defense of the spatial theory of voting as a "complete theory of voting". See e.g. Roemer (2001) and Degan and Merlo (2011) for relevant extensions and empirical applications of this framework. See Green and Shapiro (1994) for a criticism of the value of the research on party competition models in the Downsian tradition.

behavior such as acting on behalf of the interests of the members of one's social networks (see Gintis 2016). By doing so, we intended to provide an individually rational non-altruistic explanation of both turnout and voting behaviors. The first aim of this section is to substantiate the assumptions of this explanation. The second aim is to identify and discuss the limitations and possible extensions of our framework.

4.4.1 A Behavioral Voting Theory

We start by pointing out additional theoretical and empirical support for some of the social cognitive processes that underlay our model. In particular, we focus on the following assumptions: (i) party identification and social identity conflicts are relevant phenomenons, (ii) some citizens behave *as if* they have multiple party identifications, (iii) citizens with two identifications will tend to adopt a compromise behavior, and (iv) the absence of shared ideological views by the two parties induces conflicted partisans to abstain. While we mention some additional empirical evidence in support of these hypotheses, our model aims to be a *behavioral voting theory* that points towards new and open avenues for empirical research. The aim of this section is not, therefore, to substantiate that these assumptions will be *always* verified, but to suggest that they might be sensible approximations in some contexts and for some voters.

4.4.1.1 Party Identification and Social Identity Conflicts

The primary, and to some extent uncontroversial hypothesis in our model, is the long held idea that party identification is a strong influence in citizens' preferences and ultimate choices (see Campbell, Gurin and Miller 1954 and 1960 for seminal contributions). If questions concerning the stability and primacy of this effect continue to be a focus of attention and controversy, there is by now a considerable amount of evidence on its pervasive influence (e.g. Green et al. 2002; Hershey 2015)[21]. Further-

[21] On the topic of stability, see e.g. Campbell et al. (1960), Fiorina (1981), Green and Palmquist (1990), and Bartels (2002) for competing views.

more, the influence of a sense of attachment to different groups is backed by several experimental results that suggest that even induced social identities may influence behavior (e.g. Chen and Chen 2011). In the electoral context, experiments have shown that even *minimal group identification* both increases voter turnout (Schram and Sonnemans 1996; Robalo, Schram and Sonnemans 2013) and is a significant influence on how citizens make their political choices (Feddersen, Gailmard and Sandroni 2009; Bassi, Morton and Williams 2011).

In our model, citizens can identify not only with one but with several groups. If, to the best of our knowledge, there is no direct experimental evidence either supporting or discouraging this hypothesis in the political domain, recent experiments suggest that conflicting identities are relevant for other economically meaningful preferences (e.g. Benjamin, Choi and Strickland 2010; LeBoeuf, Shafir and Bayuk 2010).[22] These studies lend support to this hypothesis by showing that it is sufficient to make one identity more salient than another (the Asian or the American identity of Asian-American subjects) to trigger different behavioral responses in terms of patience (Benjamin et al. 2010) and cooperation (LeBoeuf et al. 2010). This goes in line with earlier research suggesting that the self-concept is not monolithic but multifaceted, including as many social identities as social categories one might consider relevant to oneself (e.g. Turner 1985).

4.4.1.2 Multiple Party Identifications

Though the absence of direct evidence, political behavior seems a natural instance for identity conflicts to emerge. This claim is indirectly supported by the evidence on the conflict among identifications on successive elections. For example, Niemi, Wright and Powell (1987) estimate that during the 1970s and 1980s, 20% or more

[22] Social identity has been applied in economics to explain various types of economic behavior such as gender discrimination (Akerlof and Kranton 2000), contract theory (Akerlof and Kranton 2005), and equilibrium selection in cooperative games (Chen and Chen 2011). However, most, if not all theoretical contributions have focused on identity groups that are defined with respect to a single identification. See Kirman and Teschl (2006) and Davis (2007) for discussions of Akerlof and Kranton's (2000) framework and the inclusion of multiple identities.

of Americans were *split-level identifiers*, i.e., citizens that identify with a different party at different levels of government (e.g. identify with a different party at the national and state levels). According to a national sample of campaign contributors and two national cross-section samples, these citizens were less likely to participate in politics than consistent identifiers (Niemi et al. 1987). Split-level identifiers were on average over 20% of citizens during the successive Canadian federal and provincial elections of 1974, 1979, and 1980 (Clarke and Stewart 1987), and up to 20% of the partisans of the main left party in the French presidential elections of 2007 (Perrineau 2007).[23] Bassili (1995) also finds that Canadian citizens that identify with one party but intend to vote for the candidate of a different party take more *time* than strong partisans, measured as response latencies to two questions, to express both their voting intentions and their party identifications. One possible interpretation of this result is of a *real* cognitive conflict between multiple party identifications.

Future research should investigate if conflicting identifications are relevant for voting and turnout behaviors in one election. Very often, social identifications are ingrained human responses based on a sense of belonging to several social categories (see e.g. Turner 1982). Then, we expect multiple identifications in one election to hold even in the absence of shared ideological views.[24] In this light, conflicted voters would be bi-conceptual that relate to two political worldviews, leaning either towards one view or towards the other irrespective of party polarization. Future empirical research may also shed some light on how citizens that have multiple weak identifications relate to different party ideologies and different levels of party polarization and change.

[23] See Green et al. (2002) for some evidence that multiple identifications are not common in the Italian electoral context. Uslaner (1989) reports a similar share of split-level identifiers in the 1979 Canadian elections as Clarke and Stewart (1987), but finds evidence suggesting that these citizens participated as much as consistent partisans in that elections.

[24] But not necessarily to be resilient to strong *changes* in party polarization. Still, this is not implied by the static absence of shared ideological views. See the brief discussion on the "dynamics of party identifications" in subsection 4.4.2.

4.4.1.3 Compromise Heuristic

We believe that this view has particular bite in large elections, where voters seem to rely on cues and heuristics to overcome their information shortfalls (Downs 1957; Sniderman, Brody and Tetlock 1991; Lupia and McCubbins 1998; Lau and Redlawsk 2001). The political choices that citizens are asked to make are in general complex and the information they have to reach to a decision thin (Sniderman 2000). Heuristics are "methods for arriving at satisfactory solutions with modest amounts of computation" (Simon 1990, 11), that in the electoral context allow citizens to simplify the choice itself and arrive to a rational/reasoned choice in a cost effective way (Sniderman et al. 1991; Lupia and McCubbins 1998, 2000). Party ideologies are visible and efficient cues - easy conveyable ideological stands - that provide guidance in the political landscape. The two parties provide narratives or worldviews through which citizens can understand the world of politics, and by doing so they constrain the political choices - the "choice set" - available to citizens (Sniderman 2000).

Party identification can then be seen, in part, as a *judgmental heuristic* (Sniderman 2000). It influences how citizens consider and weight different cues. In our model, this can be divided into two effects. First, party identification serves as an attention filter in the sense that citizens, when forming their preferences, focus on the cues given by their parties (e.g. read the news/opinion articles written by the elite of their party). These cues (in our model the ideological points and acceptance regions) are seen by the voters as the behavioral prescriptions of the parties for their voting behavior. Second, one's self-image (linked to one's social identity) is a personal reference point that turns the act of voting as a possibility to incur in identity gains or losses, that is, to enhance or hurt one's identity or self-concept. The more (less) compatible is a citizen's potential vote with the ideologies of the groups she identifies with, the higher the identity gains (losses) that turn out and voting may entail. Voting, in this light, is an act that can be connected with intrinsic utility gains

and losses that depend on the candidates' positions and the behavioral prescriptions of the parties.

It follows that conflicted partisans (i) focus on the cues of the two parties and (ii) derive identity gains or losses with respect to the prescriptions of the two parties. Taken together, these hypotheses imply that the bi-partisans' identity gains and losses depend upon whether the candidates' positions are or not in line with the behavioral prescriptions (ideological cues) given by the two parties. In our model this results in what can be seen as a *compromise heuristic*: A simple process of compromise based upon the cues of the two parties that replaces complex and mixed views. According to this view, conflicted voters use a cognitive shortcut to aggregate their mixed stated preferences on different issues into a moderate revealed preference that reconciles their two identifications[25].

Further experimental investigation should clarify whether voters could actually resort to such a heuristic. Several reasons can in principle underlie such behavior. For instance, a compromise heuristic could be used to simplify a difficult choice (see Lau and Redlawsk 2001), to justify a decision to oneself and to others (see Simonson 1989), or reduce the cognitive dissonance created by two conflicting affects (see Festinger 1957). All these processes seem more likely to occur whenever conflicted partisans are able to endorse a position that is a compromise and reconciles the multiple ideologies they identify with.

4.4.1.4 Conflicted Voter's Curse

This compromise is, however, not always possible in our model. This is the case whenever the two parties do not provide options which reconcile their ideological views (empty overlap region). In this situation, conflicted partisans cannot reconcile both worldviews through the act of voting. Then, it may be more difficult to justify

[25]This cognitive shortcut is different, but in many ways related, to the compromise heuristic in Simonson (1989). The author finds that an alternative tends to gain market share when it becomes a compromise or middle option in a set. Here, conflicted partisans exhibit a preference to reconcile two ideologies and endorse, whenever they can, a compromise alternative that is in between the two more salient alternatives of the political choice set.

to oneself the act of voting. It appears, as well, that a reconciliation of the potential dissonant cognitions may be harder to accomplish. Finally, the anticipation of regret may become salient when it is inevitable to betray one's identifications through the act of voting (see subsection 4.2.3). If one interprets, in the sense of Schelling (1985), such identity gains and losses as "mental consumptions", then social identity *losses loom larger than gains* whenever voting is not an opportunity to affirm one's identity or to reconcile one's multiple identifications. If parties do not provide a compromise option, conflicted voters resent both parties and are more likely to abstain.

4.4.2 Limitations and Extensions

Our model intends to be a simple theoretical framework with illustrative implications about citizens' and candidates' behaviors in large elections. It is thus a theory limited in several respects. At the same time, it is a theory prone to several extensions. We focus on (i) the treatment of the multiple issues/dimensions of the political domain, (ii) non-partisans and (iii) party loyalty, (iv) negative identifications, (v) probabilistic voting, and (vi) the dynamics of party identifications and (vii) of party ideologies.[26]

4.4.2.1 Multiple Issues/Dimensions

One limitation of our model is the restrictive nature of preferences and perceptions, in particular when considering multiple and varied issues/dimensions. For instance, the loss function in U1 implies linear costs and that all citizens share the same level of concern over all issues. Additionally, the perceptions of the parties and candidates' positions are also shared by all citizens. However, most citizens or groups attach different levels of concern to distinct issues, and citizens from different groups have different perceptions on parties and candidates' positions (see e.g. Hetherington 2008).

[26] By focusing in some limitations we certainly abstract of some other relevant limitations/extensions of the model. One immediate limitation is the number of candidates and parties; as it is increasingly the case in many countries, politics is now often dominated by more than two groups or parties. For that reason, and although the analysis may become increasingly difficult, a relevant extension would be to consider more than two candidates or parties.

A more general formulation could attach different preferences and perceptions to different groups of citizens according to their identifications, and use weighted Euclidean preferences instead of simple Euclidean preferences. We refrained from doing so in order to focus on essentials and have general analytic results for m dimensions.

Still, we believe the model to be best suitable for political contests that concentrate over few dimensions. Our results hinge on the reduction of the dimensionality of the political domain. Citizens form their preferences based on the ideological cues given by the two parties, which shrinks the multidimensional domain into a uni-dimensional bi-polar axis. This seems to be adapted to the political debate in countries in which the increase in party polarization has led to the reduction of the dimensionality of the political debate to a single party conflict dimension. Many issues once distinct are absorbed by a "left-right" conflict, and both citizens and candidates position themselves with relation to the ideologies of the two major parties. Such a reduction would be more difficult to defend in contexts where the debate is highly multi-dimensional and varied.

In such contexts, an alternative representation would be for mixed and moderate voters to adopt the view of one party on some issues and the view of the other party on other issues. This would picture these citizens as "eclectic voters", rationally endorsing a mix of policies from two different ideologies. Then, it would be possible to escape the one-dimensional feature of many political taxonomies and represent the different groups that seem to compose the political center. At the same time, the non-direct link between stated and revealed preferences should not be overlooked. For instance, heuristics and cues are commonly used by voters to arrive at their revealed preferences (Lau and Redlawsk 2001). The compromise heuristic that we have put forward in subsection 4.4.1.3 represents one way in which some groups of mixed and moderate voters may aggregate their stated preferences into a "simplified" revealed preference; and other heuristics such as *partisan stereotypes* (Lodge and Hamill 1986; Rahn 1993) may also play an important role in explaining the turnout and voting behaviors of the different groups that compose the political center. Having frameworks

(either theoretical or empirical) that incorporate citizens with different voting "strategies" can be a valuable line of future research. This seems especially relevant with respect to the different types of voters that compose the political center (see Pew Research Center 2014a for a tentative typology of the U.S. political center).

4.4.2.2 Non-partisans

When it comes to the political sphere, we concur with the authors that hold that the rational economic maximizer is not a common actor (e.g. Green and Shapiro 1994; Gintis 2016). In our framework, we have excluded this actor by imposing that each citizen identifies at least with one group. In terms of partisan affiliation, this means that we let non-partisans/independents outside of the model.

This exclusion is indirectly backed by the evidence, at least for countries such as the U.S., that citizens are influenced by party ideologies and that they in general lean towards a given party. For instance, in 2014 only around 13% of American mixed voters stated to have no leaning, which amounts to few more than the 5% of solid liberals and the 9% of solid conservatives that stated the same (Pew Research Center 2014b). Similarly, citizens that describe themselves as political independents tend to lean towards a given party (Hawkins and Nosek 2012). Even though this does not provide direct evidence of identification, it suggests that most individuals have a partisan imprint (and possibly a weak identification) in countries like the U.S.

Still, the exclusion of non-partisans can be seen as a limitation of our framework. An alternative specification of the model could include the co-existence of party/ideological cues and citizens' exogenous ideal points. I turn next to one potential way of incorporating individual exogenous policy preferences in our framework.

4.4.2.3 Party Loyalty

For partisans, a way of including individual (semi-independent) policy preferences would be to have a distribution of ideal points within the acceptance region of their

party. For conflicted partisans, the ideal points could be distributed in the center of the political domain (between the ideological points of the two parties) and party identifications would only matter when the overlap region would be non-empty. Otherwise, they would vote solely based on their exogenous ideal point. This would be a relevant extension of our model, that could lead to different implications for both citizens' and candidates' behavior.

This extension would turn our model closer to the neo-Downsian models of *party loyalty* (e.g. Adams 2001; Merrill and Adams 2001). In these models, party identification produces an *ex-post* bias on the citizens' preferences in favor of their party's nominee, in a way that leads them to vote for the candidate of their party as long as the candidate of the opposite party is not too much closer to their exogenous ideal point. The reasons underlying this *partisan bias* are in general similar to the ones we have put forward here for party or group identification. The difference is that in our model citizens have what can be seen as an *ideological bias*. That is, citizens' ideological preferences are influenced by the ideological cues of the parties they identify with irrespective of the party's nominee's identity. This means that the citizens' ideological preferences are biased *ex-ante* by their party identifications. Candidates, either from one party or the other, need to be closer than the other candidate to the citizens' bliss points if they wish to gain their support. In accordance, party identification in our model does not imply either straight ticket voting *or* partisan stereotypes; instead, it implies an ideological bias on citizens' preferences that do not allow them to vote for candidates, irrespective of their affiliation, that do not respect the prescriptions of the party/parties they identify with.

4.4.2.4 Negative Identifications

Another dimension that is absent in our model is the potential influence of negative identifications, partisan antipathies, and out-group cues to citizens' preferences. Hostility to the opposing party is likely to be a motivator for voter turnout, and can, at least in principle, affect citizens' ideological preferences. For instance, experimental

evidence from surveys in the American context suggests that out-party cues, i.e., the endorsement of a position by the elite of the party one does not identify with, may be more powerful than in-party cues in motivating value expression (Goren, Federico and Kittilson 2009) and polarization in public opinion (Nicholson 2012). This is in line with the view that party identification represents a meaningful group affiliation as long as it implies not only "positive sentiment for one's own group, but also negative sentiment toward those identifying with opposing groups" (Iyengar, Sood and Lelkes, 2012, 2). If more evidence is needed to understand the relative importance of out-party and in-party cues and their relationship, adding this feature to the model is an interesting extension.

Similarly, we do not account for "negative" behavioral prescriptions, such as not identifying with another party or not voting for a position accepted by the other party. But it seems plausible that for some groups identifying with another group or voting for a position that is accepted by the other group is as much as a betrayal as not voting for a position accepted by the group itself. While prescriptions for not voting for a position accepted by the other group could be somewhat easily introduced in our spatial framework, prescriptions for not identifying with another group could in principle be used to explain some of the dynamics of group or party identifications (see also Section 4.4.2.6 below).

4.4.2.5 Probabilistic Voting

A relevant extension is to treat the vote as a non-degenerate random variable (see e.g. Coughlin 1984). For illustration, say that we assume that the probability of a citizen to vote for a given candidate increases with the (expected) utility of voting for this candidate as long as this utility is superior to that of voting for the other candidate, otherwise the probability of voting for the former candidate is zero (see e.g. Hinich, Ledyard and Ordeshook 1973). In our framework, partisans would then vote with the *highest probability* for a candidate that adopts a position at the preferred ideological point of their party. At the same time, it would be reasonable to assume that they

would have a negligible probability of turnout for any candidate at the frontier of the acceptance region of their party. As for bi-partisans, their turnout probability would increase with lower ideological polarization, but they would continue to have a negligible probability of turnout with high degrees of polarization. Instead, an alternative specification could use probabilistic voting to address the discontinuity implicit in the citizens' turnout behavior.

Without being exhaustive, we call attention to some of the differences in electoral outcomes that would emerge with the adoption of probabilistic voting. First, and for the two types of candidates (ideological and non-ideological), candidates would be less likely to converge to the center of the political domain. This would hold since the conflicted partisans' turnout probability would be lower than that of the partisans in the case of a weak overlap. Then, candidates would be more likely to stick closer to the ideological point of one of the parties to secure the partisans' support. Second, turnout would be higher the lower the party polarization would be (given that an overlap region exists). This would be mostly driven by bi-partisans' higher levels of participation. Third, turnout and electoral equilibrium locations would depend upon the weight that citizens attach to each group. This would imply, among other things, that in the absence of an overlap region each candidate would stick to the ideological point of one party in order to secure the highest turnout probability of partisans. This contrasts with our result for ideological candidates of an unpredictable behavior in the case that parties share no ideological views.

4.4.2.6 Dynamics of Party Identifications

Another extension concerns the dynamics and evolution of citizens' party identifications. This has deserved a lot of attention in the literature, and future research could be carried out to study it within our framework. Even if slowly, party ideologies evolve over time, and partisan realignment is an important phenomenon. For instance, Milazzo et al. (2012) find evidence that citizens' decision rules are an endogenous function of parties' policy positions. In our model, the weight associated

to each identification could be endogenous to the parties' respective ideologies (positions and size of the acceptance regions). Then, conflicted partisans could lean towards one or the other party from one election to the next in response to the behavior of the parties. In principle, it would be also possible to incorporate changes in identifications (partisan realignment) due to radical shifts of party ideologies. In this perspective, partisan preferences would be seen as a learning process (Key and Cummings 1966), that change according to the citizens' perceptions of the social groups and whether they include themselves among these groups or not (see Green et al. 2002).

4.4.2.7 Dynamics of Party ideologies

Finally, our model can be used to analyze parties' best ideological strategies from one election to the other. Parties compete to win public power, and a central aspect of this competition is their effort to define the terms of political choice (Sniderman 2000). As an illustration, say that citizens' and candidates' preferences and behavior are given (e.g. from the previous election); parties could then compete on the definition of their ideologies (ideological point and acceptance region) in order for *their* candidate to win the election. We conjecture that such an analysis would predict, among other things, that the party with a greater partisan core of support (call it the *majoritarian party*) would adopt a different strategy than the *minoritarian party*. On the one hand, the best strategy of the majoritarian party would be to try to secure a win by avoiding the perception that the two parties share ideological views (avoid an overlap region). On the other hand, and in particular if conflicted partisans were pivotal, it would be in the best interest of the minoritarian party to enlarge the political choices and find shared ideological views with the majoritarian party. If successful, conflicted partisans could lean towards the candidate of the minoritarian party and he could either tie or win against the candidate of the majoritarian party. These strategies, and their success, would depend on parameters such as the stability with respect to previously adopted positions and costs of changing ideological views. A

detailed analysis of this problem would, at least in principle, connect our framework with Roemer's (2001) model of political competition in which the dynamics within the parties are focal.

4.5 Conclusion

In this chapter, we propose a unified rational theory of turnout and voting behaviors. We combine spatial voting with a social-psychology theory of group identification, and ask how the interaction between parties, candidates, and citizens affects the choices that ideological and non-ideological candidates offer to the citizens. The contribution of this chapter is two-fold: (i) to propose a new formal framework, based on party ideologies and citizens' group identifications, to study citizens' turnout and voting behaviors; and (ii) to characterize ideological and non-ideological candidates' behavior in an election with two candidates and two parties. Our model illustrates behavioral regularities of the partisans', conflicted partisans', and candidates' behavior *vis-à-vis* the ideologies of parties or other groups.

The theoretical treatment of the behavior of conflicted voters is one of the main novelties of this chapter, which opens several empirical demanding questions. Given the difficulty of disentangling different underlying reasons behind citizens' and candidates' behaviors, one potential fruitful way to test these implications is through laboratory or field experiments.[27] One could then test, among other things, different strengths of betrayal aversion for subjects with multiple identifications.

In any case, our results remain suggestive, and by no means do we wish to claim that all the predictions (neither all the assumptions) will be empirically verified. Instead, these predictions provide, in our view, meaningful insights (illustrations) of some potential relationships between party polarization, citizens' and candidates' behavior in elections where two groups play a fundamental role in determining their

[27] See Palfrey (2009) and Morton and Williams (2008) for defenses and reviews of this line of research in political economy.

behavior. Seen in this light, it is possible to summarize the main implications of our model in a few claims.

Claim 1. *If there exist no shared ideological views then conflicted voters are more likely to abstain.*

This may capture the rational abstention of citizens that feel dissatisfied with all the possible positions available in the political domain due to excessive party polarization. While we do not expect, as our model would predict, that *all* bi-partisans would abstain in such a case, we find it useful to have such a yardstick for future research on the contextual determinants that may increase or lower such abstention. Furthermore, the conflicted voter's curse rationalizes how a group of citizens may abstain even if they could, as a group, change the outcome of the election in favor of their preferred outcome.

Claim 2. *If there exist shared ideological views and conflicted voters are pivotal then candidates tend to converge to the center of the political domain.*

This highlights sufficient conditions for candidates to converge to a position that is in between the two parties' ideologies. If these conditions hold, our model predicts that both non-ideological and ideological candidates converge to the position of the "median voter". Furthermore, our results pinpoint two reference positions in the center of the political domain in which this median voter may be positioned *ex-post*.

Claims 1 and 2 also expose why politics attracting the median voter might work better in times of lower political conflict. In these times, conflicted voters are more likely to support a candidate that tries to conciliate the narratives of both parties. It would be interesting to see if the same would hold for the other groups of voters that may compose the political center. In times of greater perceived polarization, our model suggests that it is instead rational for candidates to focus on their partisan core of support.

Claim 3. *If there exist shared ideological views and conflicted voters are not pivotal then majoritarian ideological candidates can adopt a progressive behavior while minoritarian ideological candidates tend to adopt a position at the center of the political domain.*

This means that if party polarization is weak and there are few conflicted partisans then majoritarian candidates have strategic flexibility to adopt any position around the preferred ideology of their party. At the same time, it is rational for minoritarian candidates to adopt a moderate position at the center of the political domain in order to attract conflicted partisans to turn out and vote for them.

Claim 4. *If there exist no shared ideological views then ideological candidates can adopt any position around their party's ideology.*

This indicates that candidates' behavior may be quite unpredictable in cases of strong perceived polarization in party ideologies in our setting. Claims 3 and 4 illustrate a potential relationship between party polarization, ideological coherence, and the possibility of candidates to take progressive or more extreme positions. These claims also highlight the possible tension between the loyalty to a party (or ideological credibility) and the will to maximize plurality.

The identification to parties or other identity groups, as a source of an ideological bias, is one possibility to reconcile turnout and voting behaviors with other types of social-cooperative behaviors. Only by resorting to behavioral theories that explain the origin, strength, and extent of the underlying reasons of these behaviors can we hope to have *ex-ante* rationalizations of the act of voting (Mueller 2003). Our model is a modest attempt at this, which we hope can encourage further research on multiple party identifications, ideological bias, and the behavior of the voters that compose the political center.

Bibliography

Adams, J. (2001) A Theory of Spatial Competition with Biased Voters: Party Policies Viewed Temporally and Comparatevely. British Journal of Political Science 31: 121–58.

Akerlof, G. A. and R. E. Kranton (2000) Economics and Identity. The Quarterly Journal of Economics 115(3): 715–53.

——— (2005) Identity and the Economics of Organizations. The Journal of Economic Perspectives 19(1): 9–32.

Aldrich, J. H. (1983) A Downsian Spatial Model with Party Activism. The American Political Science Review 77(4): 974–90.

——— (1993) Rational Choice and Turnout. American Journal of Political Science 37(1): 246–78.

Aldrich, J. H. and M. D. McGinnis (1989) A Model Of Party Constraints on Optimal Candidate Positions. Mathematical and Computer Modelling 12(4/5): 437–50.

Aranson, P. H., M. J. Hinich, and P. C. Ordeshook (1974) Election Goals and Strategies: Equivalent and Nonequivalent Candidate Objectives. The American Political Science Review 68(1): 135–52.

Bartels, L. M. (2002) Beyond the Running Tally: Partisan Bias in Political Perceptions. Political Behavior 24(2): 117–50.

Bassi, A., R. B. Morton, and K. C. Williams (2011) The Effects of Identities, Incentives, and Information on Voting. The Journal of Politics 73(2): 558–71.

Bassili, J. N. (1995) On the Psychological Reality of Party Identification: Evidence from the Accessibility of Voting Intentions and of Partisan Feelings. Political Behavior 17(4): 339–58.

Benjamin, D. J., J. J. Choi, and A. J. Strickland (2010) Social Identity and Preferences. The American Economic Review 100(4): 1913–28.

Black, D. (1958) The Theory of Committees and Elections. Cambridge University Press., Cambridge.

Bowman, D., D. Minehart, and M. Rabin (1999) Loss Aversion in a Consumption-Savings Model. Journal of Economic Behavior and Organization 38: 155–78.

Brennan, G. and A. Hamlin (1998) Expressive Voting and Electoral Equilibrium. Public Choice 95: 149–75.

Budge, I., D. Robertson, and D. Hearl (1987) Ideology, Strategy, and Party Change: Spatial Analyses of Post-war Election Programmes in 19 Democracies. Cambridge University Press, Cambridge.

Campbell, A., P. E. Converse, W. E. Miller, and D. E. Stokes (1960) The American Voter. Wiley, New York.

Campbell, A., G. Gurin, and W. E. Miller (1954) The Voter Decides. Row, Peterson, Evanston, IL.

Carsey, T. and G. Laymain (2006) Changing Sides or Changing Minds? Party Identification and Policy Preferences in the American Electorate. American Journal of Political Science 50(2): 464–77.

Chen, R. and Y. Chen (2011) The Potential of Social Identity for Equilibrium Selection. The American Economic Review 101(6): 2562–89.

Clarke, H. D. and M. C. Stewart (1987) Partisan Inconsistency and Partisan Change in Federal States: The Case of Canada. American Journal of Political Science 31(2): 383–407.

Coughlin, P. J. (1984) Davis-Hinich Conditions and Median Outcomes in Probabilistic Voting Models. Journal of Economic Theory 34: 1–12.

Coughlin, P. J. and S. Nitzan (1981) Directional and Local Electoral Equilibria with Probabilistic Voting. Journal of Economic Theory 24(2): 226–39.

Crockett, M. J., Z. Kurth-Nelson, J. Z. Siegel, P. Dayan, and R. J. Dolan (2014) Harm to Others Outweighs Harm to Self in Moral Decision Making. Proceedings of the National Academy of Sciences 111(48): 17320–25.

Dancey, L. and P. Goren (2010) Party Identification, Issue Attitudes, and the Dynamics of Political Debate. American Journal of Political Science 54(3): 686–99.

Davis, J. B. (2007) Akerlof and Kranton on Identity in Economics: Inverting the Analysis. Cambridge Journal of Economics 31: 349–62.

Degan, A. and A. Merlo (2011) A Structural Model of Turnout and Voting in Multiple Elections. Journal of the European Economic Association 9(2): 209–45.

Dhar, R. and K. Wertenbroch (2000) Consumer Choice Between Hedonic and Utilitarian Goods. Journal of Marketing Research 37(1): 60–71.

Dhillon, A. and S. Peralta (2002) Economic Theories of Voter Turnout. The Economic Journal 112: 332–52.

Downs, A. (1957) An Economic Theory of Democracy. Harper, New York.

Enelow, J. M. and M. J. Hinich (1984) The Spatial Theory of Voting. Cambridge University Press, Cambridge.

Evans, G. and R. Andersen (2004) Do Issues Decide? Partisan Conditioning and Perceptions of Party Issue Positions across the Electoral Cycle. Journal of Elections, Public Opinion and Parties 14(1): 18–39.

Feddersen, T. (2004) Rational Choice Theory and the Paradox of Not Voting. The Journal of Economic Perspectives 18(1): 99–112.

Feddersen, T., S. Gailmard, and A. Sandroni (2009) Moral Bias in Large Elections: Theory and Experimental Evidence. American Political Science Review 103(2): 175–92.

Festinger, L. A. (1957) A Theory of Cognitive Dissonance. Row, Peterson, Evanston, IL.

Fiorina, M. P. (1981) Retrospective Voting in American National Elections. Yale University Press, New Haven.

Fiorina, M. P. and S. J. Abrams (2008) Polarization in the American Public. Annual Review of Political Science 11: 563–88.

Fiorina, M. P., S. J. Abrams, and J. C. Pope (2006) Culture War? The Myth of a Polarized America. Longman, New York.

——— (2008) Polarization in the American Public: Misconceptions and Misreadings. The Journal of Politics 70(2): 556–60.

Flinn, T. A. and F. M. Wirt (1965) Local Party Leaders: Groups of Like Minded Men. Midwest Journal of Political Science 9: 77–98.

Galanter, E. and P. Pliner (1974) Cross-Modality Matching of Money Against Other Continua. In: H. R. Moskowitz, B. Scharf, and J. C. Stevens (eds) Sensation and Measurement. D. Reidel Publishing Company, Dordrecht, Holland: 65–76.

Geys, B. (2006) "Rational" Theories of Voter Turnout: A Review. Political Studies Review 4: 16–35.

Gintis, H. (2016) Homo Ludens: Social Rationality and Political Behavior. Journal of Economic Behavior and Organization 126: 95–109.

Goren, P. (2005) Party Identification and Core Political Values. American Journal of Political Science 49(4): 881–96.

Goren, P., C. M. Federico, and M. C. Kittilson (2009) Source Cues, Partisan Identities, and Political Value Expression. American Journal of Political Science 53(4): 805–20.

Green, D. P. and B. Palmquist (1990) Of Artifacts and Partisan Instability. American Journal of Political Science 34: 872–902.

Green, D. P., B. Palmquist, and E. Schickler (2002) Partisan Hearts and Minds: Political Parties and the Social Identities of Voters. Yale ISPS series, Yale University Press, New Haven.

Green, D. P. and I. Shapiro (1994) Pathologies of Rational Choice Theory: A Critique of Applications in Political Science. Yale University Press, New Haven.

Grofman, B. (2004) Downs and Two-Party Convergence. Annual Review of Political Science 7: 25–46.

Hawkins, C. B. and B. A. Nosek (2012) Motivated Independence? Implicit Party Identity Predicts Political Judgments among Self-proclaimed Independents. Personality and Social Psychology Bulletin 38(11): 1437–52.

Hershey, M. R. (2015) Party politics in America. Pearson Education, New York,16th edition.

Hetherington, M. J. (2008) Turned off or Turned on? How Polarization Affects Political Engagement. In: P. Nivola and D. Brady (eds) Red and Blue Nation. 2 Brookings Institution Press and Hoover Institution, Washington, DC, and Stanford, CA: 1–33.

Hinich, M. J., J. O. Ledyard, and P. C. Ordeshook (1972) Nonvoting and the Existence of Equilibrium under Majority Rule. Journal of Economic Theory 4: 144–53.

——— (1973) A Theory of Electoral Equilibrium: A Spatial Analysis Based on the Theory of Games. The Journal of Politics 35(1): 154–93.

Hinich, M. J. and P. C. Ordeshook (1969) Abstentions and Equilibrium in the Electoral Process. Public Choice 7: 76–106.

Hotelling, H. (1929) Stability in Competition. The Economic Journal 39: 41–57.

Iyengar, S., G. Sood, and Y. Lelkes (2012) Affect, Not Ideology A Social Identity Perspective on Polarization. Public Opinion Quarterly 76(3): 405–31.

Kahneman, D. and A. Tversky (1982) The Simulation Heuristic. In: D. Kahneman, P. Slovic, and A. Tversky (eds) Judgement under uncertainty: Heuristics and Biases. Cambridge University Press.

Katz, R. S. (1979) The Dimensionality of Party Identification: Cross-National Perspectives. Comparative Politics 11(2): 147–63.

Key, V. O. and M. C. Cummings (1966) The Responsible Electorate. Belknap Press of Harvard University Press.

Kirman, A. and M. Teschl (2006) Searching for Identity in the Capability Space. Journal of Economic Methodology 13(3): 299–325.

Kramer, G. H. and A. K. Klevorick (1974) Existence of a "Local" Co-operative Equilibrium in a Class of Voting Games. The Review of Economic Studies 41(4): 543–47.

Lau, R. R. and D. P. Redlawsk (2001) Advantages and Disadvantages of Cognitive Heuristics in Political Decision Making. American Journal of Political Science 45(4): 951–71.

LeBoeuf, R. A., E. Shafir, and J. B. Bayuk (2010) The Conflicting Choices of Alternating Selves. Organizational Behavior and Human Decision Processes 111(1): 48–61.

Levendusky, M. (2009) The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans. University of Chicago Press, Chicago.

Lodge, M. and R. Hamill (1986) A Partisan Schema for Political Information Processing. The American Political Science Review 80(2): 505–20.

Lupia, A. and M. D. McCubbins (1998) The Democratic Dilemma. Can Citizens Learn What They Need to Know?. Cambridge University Press, Cambridge.

——— (2000) The Institutional Foundations of Political Competence: How Citizens Learn What They Need to Know. In: A. Lupia, M. D. McCubbins, and S. L. Popkin (eds) Elements of Reason: Cognition, Choice, and the Bounds of Rationality. Cambridge University Press, Cambridge: 47–66.

Matthews, S. A. (1979) A Simple Direction Model of Electoral Competition. Public Choice 34(2): 141–56.

Merrill, S. and J. Adams (2001) Computing Nash equilibria in Probabilistic, Multiparty Spatial Models with Nonpolicy Components. Political Analysis 9(4): 347–61.

Milazzo, C., J. Adams, and J. Green (2012) Are Voter Decision Rules Endogenous to Parties' Policy Strategies? A Model with Applications to Elite Depolarization in Post-Thatcher Britain. The Journal of Politics 74(1): 262–76.

Morton, R. B. (1987) A Group Majortiy Voting Model of Public Good Provision. Social Choice and Welfare 4: 117–31.

——— (1991) Groups in Rational Turnout Models. American Journal of Political Science 35(3): 758–76.

Morton, R. B. and K. C. Williams (2008) Experimentation in Political Science. The Oxford Handbook of Political Methodology: 339–56.

Mueller, D. C. (2003) Public Choice III. Cambridge Universiy Press, Cambridge.

Nicholson, S. P. (2012) Polarizing Cues. American Journal of Political Science 56(1): 52–66.

Niemi, R. G., S. Wright, and L. W. Powell (1987) Multiple Party Identifiers and the Measurement of Party Identification. The Journal of Politics 49(4): 1093–103.

Palfrey, T. R. (2009) Laboratory Experiments in Political Economy. Annual Review of Political Science 12: 379–88.

Perrineau, P. (2007) Électeurs Dissonants et Électeurs Fidèles. Revue Française de Sciense Politique 57: 343–52.

Pew Research Center (2014a) Beyond Red vs. Blue: The Political Typology. http://www.people-press.org/2014/06/26/the-political-typology-beyond-red-vs-blue/.

——— (2014b) Political Polarization in the American Public. http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/.

Rahn, W. M. (1993) The Role of Partisan Stereotypes in Information Processing about Political Candidates. American Journal of Political Science 37(2): 472–96.

Riker, W. H. and P. C. Ordeshook (1968) A Theory of the Calculus of Voting. American Political Schence Review 62: 25–42.

Robalo, P., A. Schram, and J. Sonnemans (2013) Other-regarding Preferences, Group Identity and Political Participation: An Experiment. Tinbergen Institute Discussion Paper.

Roemer, J. E. (2001) Political Competition. Harvard University Press, Cambridge, MA.

Samuelson, L. (1984) Electoral Equilibria with Restricted Strategies. Public Choice 43(3): 307–27.

Schelling, T. C. (1985) The Mind as a Consuming Organ. In: J. Elster (ed) The Multiple Self. Cambridge University Press, Cambridge.

Schram, A. and J. Sonnemans (1996) Why People Vote: Experimental Evidence. Journal of Economic Psychology 17: 417–42.

Shayo, M. (2009) A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution. American Political Science Review 103(2): 147–74.

Shayo, M. and A. Harel (2012) Non-consequentialist Voting. Journal of Economic Behavior and Organization 81: 299–313.

Simon, H. A. (1990) Invariants of Human Behavior. Annual Review of Psychology 41: 1–19.

Simonson, I. (1989) Choice Based on Reasons: The Case of Attraction and Compromise Effects. Journal of Consumer Research 16(2): 158–74.

Sniderman, P. M. (2000) Taking Sides: A Fixed Choice Theory of Political Reasoning. In: A. Lupia, M. D. McCubbins, and S. L. Popkin (eds) Elements of Reason: Cognition, Choice, and the Bounds of Rationality. Cambridge University Press, Cambridge: 67–75.

Sniderman, P. M., R. A. Brody, and P. Tetlock (1991) Reasoning and Choice: Explorations in Political Psychology. Cambridge University Press, Cambridge.

Turner, J. C. (1982) Towards a Cognitive Redefinition of the Social Group. In: H. Tajfel (ed) Social Identity and Intergroup Relations. Cambridge University Press, New York.

——— (1985) Social Categorization and the Self-concept: A Social Cognitive Theory of Group Behavior. In: E. J. Lawler (ed) Advances in Group Processes. 2 JAI Press, Greenwich, CT: 77–121.

Tversky, A. and D. Kahneman (1991) Loss Aversion in Riskless Choice: A Reference-dependent Model. The Quarterly Journal of Economics 107(4): 1039–1061.

Uhlaner, C. J. (1989) "Relational Goods" and Participation: Incorporating Sociability into a Theory of Rational Action. Public Choice 62: 253–85.

Uslaner, E. M. (1989) Multiple Party Identifiers in Canada: Participation and Affect. The Journal of Politics 51(4): 993–1003.

Vendrik, M. C. M. and G. B. Woltjer (2007) Happiness and Loss Aversion: Is Utility Concave or Convex in Relative Income. Journal of Public Economics 91: 1423–48.

Weisberg, H. F. (1980) A Multidimensional Conceptualization of Party Identification. Political Behavior 2(1): 33–60.

Appendix

PROOF OF PROPOSITION 1:

Proof. For any $i \in N$, suppose that it does not exist $k \in \{1,2\}$ such that $||\delta_p - \theta_k|| \leq d_p$ for all $p \in P_i$. It follows from (4.1)-(4.3) that for any θ_k we have $u_i(\theta_k) = \sum_{p \in P_i}(-\frac{I_i^p}{d_p}||\delta_p - x||) + I_i^q - c_i$ if $||\delta_q - x|| \leq d_q$ for only one $q \in P_i$ and $\#P_i = 2$ or $u_i(\theta_k) = \sum_{p \in P_i}(-\frac{I_i^p}{d_p}||\delta_p - x||) - c_i$ otherwise. Since $c_i > I_i^p$ for any $p \in P_i$, it follows that $u_i(\theta_k) < 0$ for any $k \in \{1,2\}$. If $t_i = 0$, then for any θ_k we have $t_i[u_i(\theta_k)] = 0$. Thus $t_i = 0$ maximizes utility. Now suppose there exists $k \in \{1,2\}$ such that $||\delta_p - \theta_k|| \leq d_p$ for all $p \in P_i$. It follows from (4.1)-(4.3) that there exists $k \in \{1,2\}$ such that $u_i(\theta_k) = \sum_{p \in P_i}(I_i^p - \frac{I_i^p}{d_p}||\delta_p - x||)$. If $t_i = 1$, then $t_i[u_i(\theta_k)] \geq 0$ for some $k \in \{1,2\}$. Therefore $t_i = 1$ maximizes utility.

□

PROOF OF LEMMA 1:

Proof. (a.1 and a.2) Suppose $i \in N_b$. From (4.1) we have that $u_i(\delta_b) > u_i(x)$ for all $x \in X$ such that $\delta_b \neq x$, since any deviation from δ_b to x entails $u_i(x) - u_i(\delta_b) = -\frac{I_i^b}{d_b}||\delta_b - x|| < 0$ if $x \in A_b$ and $u_i(x) - u_i(\delta_b) = -(\frac{I_i^b}{d_b}||\delta_b - x|| + I_i^b + c_i) < 0$ otherwise. Given Proposition 1, it follows that for all $i \in N_b$ there always exists $x_i^* \in X$ such that $x_i^* = \underset{t_i \in \{0,1\},\ x \in X}{\operatorname{argmax}} t_i[u_i(x)] = \delta_b$. An analogous reasoning proves that for all $i \in N_r$ there always exists $x_i^* \in X$ such that $x_i^* = \underset{t_i \in \{0,1\},\ x \in X}{\operatorname{argmax}} t_i[u_i(x)] = \delta_r$.

(b.1 and b.2) Suppose $i \in N_{br}$. Since $\frac{I_i^b}{d_b} \neq \frac{I_i^r}{d_r}$ for any $i \in N_{br}$, then either $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$ or $\frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}$ for all $i \in N_{br}$. Assume $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$, and that $O_{br} \neq \emptyset$. From (4.1) we have that $u_i(\alpha_b) > u_i(x)$ for all $x \in X$ such that $\alpha_b \neq x$, since any deviation from α_b to x entails

$u_i(x) - u_i(\alpha_b) = \frac{I_i^b}{d_b}||\delta_b - x|| + \frac{I_i^r}{d_r}||\delta_r - x|| - (\frac{I_i^b}{d_b}||\delta_b - \alpha_b|| + \frac{I_i^r}{d_r}||\delta_r - \alpha_r||) < 0$ if $x \in O_{br}$, $u_i(x) - u_i(\alpha_b) = \frac{I_i^b}{d_b}||\delta_b - x|| + \frac{I_i^r}{d_r}||\delta_r - x|| - (\frac{I_i^b}{d_b}||\delta_b - \alpha_b|| + \frac{I_i^r}{d_r}||\delta_r - \alpha_b|| + I_i^b + c_i) < 0$ if $x \in A_r \backslash \{O_{br}\}$, $u_i(x) - u_i(\alpha_b) = \frac{I_i^b}{d_b}||\delta_b - x|| + \frac{I_i^r}{d_r}||\delta_r - x|| - (\frac{I_i^b}{d_b}||\delta_b - \alpha_b|| + \frac{I_i^r}{d_r}||\delta_r - \alpha_b|| + I_i^r + c_i) < 0$ if $x \in A_b \backslash \{O_{br}\}$, and $u_i(x) - u_i(\alpha_b) = \frac{I_i^b}{d_b}||\delta_b - x|| + \frac{I_i^r}{d_r}||\delta_r - x|| - (\frac{I_i^b}{d_b}||\delta_b - \alpha_b|| + \frac{I_i^r}{d_r}||\delta_r - \alpha_b|| + \frac{I_i^b}{d_b} + \frac{I_i^r}{d_r} + c_i) < 0$ if $x \notin A_b \cup A_r$. It follows that for all $i \in N_{br}$ such that $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$, if $O_{br} \neq \emptyset$ then there exists $x_i^* \in X$ such that $x_i^* = \underset{t_i \in \{0,1\},\ x \in X}{\operatorname{argmax}} t_i[u_i(x)] = \alpha_b$. An analogous reasoning proves that for all $i \in N_{br}$ such that $\frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}$ there exists $x_i^* \in X$ such that $x_i^* = \underset{t_i \in \{0,1\},\ x \in X}{\operatorname{argmax}} t_i[u_i(x)] = \alpha_r$ if $O_{br} \neq \emptyset$. Now suppose $O_{br} = \emptyset$. It follows that for all $i \in N_{br}$ it does not exist $x \in X$ such that $||\delta_p - x|| \leqslant d_p$ for all $p \in P_i$. Then, Proposition 1 implies that for all $i \in N_{br}$ it does not exist $x \in X$ such that $x = \underset{t_i \in \{0,1\},\ x \in X}{\operatorname{argmax}} t_i[u_i(x)]$.

□

PROOF OF THEOREM 1:

Proof. Since $O_{br} \neq \emptyset$, it follows from Lemma 1 that for all $i \in N$ there exists $x_i^* \in X$ such that $x_i^* = \underset{t_i \in \{0,1\},\ x \in X}{\operatorname{argmax}} t_i[u_i(x)]$ and $x_i^* \in \{\delta_b, \delta_r, \alpha_b, \alpha_r\}$. From P1-P2 and (4.1) we have that δ_b, δ_r, α_b, and α_r can all be connected with a line segment $L = \{(1-l)\delta_b + l\delta_r : l \in R^{++}\}$. For any $x \in L$, let N_R^x denote the number of x_i^* that are on one side of the line with respect to x including the ones that lie on x and N_L^x denote the number of x_i^* that are on the other side of the line with respect to x including the ones that lie on x. Then, since N is odd, there exists $x \in \{\delta_b, \delta_r, \alpha_b, \alpha_r\}$ such that $N_R^x \geq \frac{N}{2}$ and $N_L^x \leq \frac{N}{2}$. This point guarantees a tie while any unilateral deviation from this point either entails a loss or no strict advantage. It follows that there exists a convergent equilibrium $(\theta_1^*, \theta_2^*) = (\theta^*, \theta^*)$ at one of the strategies $\theta^* \in \{\delta_b, \delta_r, \alpha_b, \alpha_r\}$.

□

PROOF OF THEOREM 2:

Proof. Given Theorem 1, there are two remaining cases in which $O_{br} \neq \emptyset$ and B1 holds. The first is when either $N_b = N_r + N_{br}$ or $N_r = N_b + N_{br}$. Suppose that $N_b = N_r + N_{br}$.

It follows from $N_b = N_r + N_{br}$ and $N_b + N_r + N_{br} = N$ that $N_b = \frac{N}{2}$ and $N_r + N_{br} = \frac{N}{2}$. Lemma 1, A1, and B1 imply that there is at least one $i \in N_{br}$ such that $x_i^* = \alpha_b$. Then δ_b and α_b guarantee a tie against each other while any unilateral deviation from one of these points either entails a loss or no strict advantage. This is the case since any $x \in A_b \backslash \{O_{br}\}$ looses against δ_b and α_b beats any $x \in A_r$ such that $x \neq \alpha_b$: α_b guarantees at least $N_b + 1 = \frac{N}{2} + 1$ votes against any distinct $x \in A_r$. Thus, while δ_b and α_b guarantee a tie against each other any deviation from one of these points does not guarantee a tie. An analogous reasoning proves that if $N_r = N_b + N_{br}$ then δ_r and α_r guarantee a tie against each other while any deviation from one of these points does not. Therefore there exist four equilibrium pairs (θ_1^*, θ_2^*) such that $\theta_1^*, \theta_2^* \in \{\delta_b, \alpha_b\}$ if $N_b = N_r + N_{br}$ and $\theta_1^*, \theta_2^* \in \{\delta_r, \alpha_r\}$ if $N_r = N_b + N_{br}$.

The second case is when N is even and neither $N_b = N_r + N_{br}$ nor $N_r = N_b + N_{br}$. Given B1, either $N_b + \#\{i \in N_{br} : \frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}\} > N_r + \#\{i \in N_{br} : \frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}\}$ or $N_b + \#\{i \in N_{br} : \frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}\} < N_r + \#\{i \in N_{br} : \frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}\}$. It is straightforward to see that there exists a convergent equilibrium $(\theta_1^*, \theta_2^*) = (\theta^*, \theta^*)$ at one of the strategies $\theta^* \in \{\delta_b, \delta_r, \alpha_b, \alpha_r\}$.

Now suppose that B1 does not hold, i.e., that $N_b + \#\{i \in N_{br} : \frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}\} = N_r + \#\{i \in N_{br} : \frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}\}$. Assume there are $i, j \in N_{br}$ such that $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$ and $\frac{I_j^b}{d_b} < \frac{I_j^r}{d_r}$. It follows that any point within O_{br} that is on the line segment that connects δ_b to δ_r can guarantee a tie against any other of these points. This implies that all the possible combinations of these points are equilibrium pairs. Now suppose that are no $i, j \in N_{br}$ such that $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$ and $\frac{I_j^b}{d_b} < \frac{I_j^r}{d_r}$. Then, A1 and $N_b + \#\{i \in N_{br} : \frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}\} = N_r + \#\{i \in N_{br} : \frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}\}$ imply that either $N_b = N_r + \#\{i \in N_{br} : \frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}\}$ or $N_r = N_b + \#\{i \in N_{br} : \frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}\}$. If $N_b = N_r + \#\{i \in N_{br} : \frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}\}$, then δ_b and any point within O_{br} that is on the line segment that connects δ_b to δ_r can guarantee a tie against any other of these points. The analogous holds if $N_r = N_b + \#\{i \in N_{br} : \frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}\}$. In both cases all possible combinations of these points are equilibrium pairs, i.e., there exists a continuum of possible equilibrium pairs.

□

PROOF OF THEOREM 3:

Proof. Since $O_{br} = \emptyset$, it follows from Lemma 1 that for all $i \in N_{br}$ it does not exist $x \in X$ such that $x = \underset{t_i \in \{0,1\},\ x \in X}{\operatorname{argmax}} t_i[u_i(x)]$ (i.e., $t_i = 0$ maximizes utility; see also Proposition 1). Lemma 1 also implies that $x_i^* = \delta_b$ for all $i \in N_b$ and $x_i^* = \delta_r$ for all $i \in N_r$. Suppose that $N_b \neq N_r$. Then, either $N_b > N_r$ or $N_b < N_r$. Assume $N_b > N_r$. Then δ_b guarantees a tie while any unilateral deviation from this point either entails a loss or no strict advantage. The analogous holds if $N_b < N_r$. It follows that there exists a convergent equilibrium $(\theta_1^*, \theta_2^*) = (\theta^*, \theta^*)$ at $\theta^* = \delta_b$ if $N_b > N_r$ and at $\theta^* = \delta_r$ if $N_b < N_r$.

Now suppose that $N_b = N_r$. Then, both δ_b and δ_r guarantee a tie against each other while any unilateral deviation from one of these points either entails a loss or no strict advantage. Thus, there exist four equilibrium pairs (θ_1^*, θ_2^*) such that $\theta_1^*, \theta_2^* \in \{\delta_b, \delta_r\}$.

□

PROOF OF THEOREM 4:

Proof. From C1 and N odd, we have $N_{br} > |N_b - N_r|$. Lemma 1 and $O_{br} \neq \emptyset$ imply that for all $i \in N$ there exists $x_i^* \in X$ such that $x_i^* = \underset{t_i \in \{0,1\},\ x \in X}{\operatorname{argmax}} t_i[u_i(x)]$ and $x_i^* \in \{\delta_b, \delta_r, \alpha_b, \alpha_r\}$, and from P1-P2 and (4.1) we have that δ_b, δ_r, α_b, and α_r can all be connected with a line segment $L = \{(1-l)\delta_b + l\delta_r : l \in R^{++}\}$. Since N is odd there exists $x \in L$ such that $N_R^x \geq \frac{N}{2}$ and $N_L^x \leq \frac{N}{2}$ (see proof of Theorem 1). Then, since $N_{br} > |N_b - N_r|$, this point is a convergent equilibrium $(\theta_1^*, \theta_2^*) = (\theta^*, \theta^*)$ at one of the strategies $\theta^* \in \{\alpha_b, \alpha_r\}$.

Suppose now that N is even. Then, it is possible that $N_b + \#\{i \in N_{br} : \frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}\} = N_r + \#\{i \in N_{br} : \frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}\}$ (and $N_{br} = |N_b - N_r|$ and $O_{br} \neq \emptyset$). In that case, a convergent equilibrium is not the solution of the electoral competition game (see proof of Theorem 2).

Now suppose that C1 does not hold, i.e., $N_{br} < |N_b - N_r|$. Then either $N_b > N_r + N_{br}$ or $N_r > N_b + N_{br}$. Assume $N_b > N_r + N_{br}$, and that there are $i, j \in N_{br}$ such that $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$ and $\frac{I_j^b}{d_b} < \frac{I_j^r}{d_r}$. Given that $O_{br} \neq \emptyset$, C2, and $X_2 = A_r$, the closest position to δ_b that

candidate 2 can adopt is α_b. Then, since $N_b > N_r + N_{br}$, candidate 1 guarantees a win and the same maximal plurality at any $x \in A_b$ that is at a lower distance from δ_b than α_b. Candidate 2 maximizes plurality at α_b, since α_b guarantees at least $N_b + 1 = \frac{N}{2} + 1$ votes against any $x \in A_r$ and $N_r + N_{br}$ against the dominant positions of candidate 1. Any unilateral deviation from these points entails no strict advantage or a decrease in plurality. All possible combinations of these points are equilibrium pairs: There exist a continuum of equilibrium pairs (θ_1^*, θ_2^*) such that $\theta_1^* \in ||\delta_b - \theta_1^*|| < ||\delta_b - \alpha_b||$ and $\theta_2^* = \alpha_b$. An analogous result holds if there is no $i \in N_{br}$ such that $\frac{I_i^b}{d_b} < \frac{I_i^r}{d_r}$ (but there exists $i \in N_{br}$ such that $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$). Now suppose there is no $i \in N_{br}$ such that $\frac{I_i^b}{d_b} > \frac{I_i^r}{d_r}$. Then, candidate 2 guarantees the same maximal plurality at any point within O_{br} that is on the line segment that connects δ_b to δ_r, while candidate 1 guarantees the same maximal plurality at the same set of points as before. All possible combinations of these points are equilibrium pairs. If we assume instead that $N_r > N_b + N_{br}$, then analogous results to these hold.

Suppose now that $O_{br} = \emptyset$. Given that $X_1 = A_b$ and $X_2 = A_r$, candidate 1 cannot adopt any position $x \in A_r$ and candidate 2 cannot adopt any position $x \in A_b$. It follows that candidate 1 guarantees the same maximal plurality at any $x \in A_b$ and candidate 2 guarantees the same maximal plurality at any $x \in A_r$. All possible combinations of these points are equilibrium pairs: There exist a continuum of equilibrium pairs (θ_1^*, θ_2^*) such that $\theta_1^* \in A_b$ and $\theta_2^* \in A_r$.

□

PROOF OF THEOREM 5:

Proof. See proof of Theorem 4. □

PROOF OF THEOREM 6:

Proof. See proof of Theorem 4. □

Chapter 5

On the Roots of the Intrinsic Value of Decision Rights: Evidence from France and Japan*

Bartling, Fehr and Herz (2014) found that Swiss individuals attach an economically meaningful intrinsic value to make a decision by themselves rather than delegating it to another person. We refine their analysis in order to disentangle how much of such value stems from (i) a preference for independence from others, (ii) a desire for power, or (iii) other motives such as a preference for self-reliance, and conduct a cross-cultural comparison between France and Japan. Our findings suggest that (i) Japanese and French individuals intrinsically value decision rights beyond their instrumental benefit, that (ii) this positive value is greater for French than Japanese individuals, and that (iii) self-reliance is the only rationale behind the intrinsic value of decision rights in both France and Japan. These results bring new insights into the roots of the preference for being in control, which can be relevant for institutional design.

Keywords: Intrinsic value; Decision rights; Independence; Power; Self-reliance; Cross-cultural experiment.

*This chapter is adapted from a joint work with Nobuyuki Hanaki and Benoît Tarroux.

5.1 Introduction

Who holds the control over decisions is an important dimension of many social interactions. In organizations, markets, or civil society, individuals interact by keeping or delegating the control over diverse decisions. In economics, the principal-agent dilemma over decision rights is the focus of a prolific literature (e.g. Simon 1951; Hart and Moore 1990; Aghion and Tirole 1997). While most authors have looked at decision rights as *instrumental* to achieve certain outcomes, holding the control over decisions may be valued for its own sake, i.e., even when it carries no advantage in terms of achieving preferred outcomes. Indeed, in a recent experimental study that separates the intrinsic and instrumental values of decision rights on the basis of subjects' revealed preferences, Bartling et al. (2014) provide some evidence that individuals (in their context Swiss students) attach an economically meaningful *intrinsic* value to make a decision for themselves rather than delegating it to another person.

A natural question that follows is what are the possible *rationales* behind such an intrinsic value of decision rights. One potential underlying reason is a preference for autonomy on the process of choice. If this is the case, decision rights carry an intrinsic value beyond their instrumental value either due to a desire to implement one's decision (a sense of *self-reliance*) or a preference for *independence* from the interference of another person. An alternative reason is a preference for *power* associated with holding the decision right. In this case, the intrinsic value of holding control stems from a desire to be able to decide on behalf of someone else.

While both autonomy and power are reasonable sources of an intrinsic value of decision rights, they are considerably different in nature and implications. On the one hand, self-reliance and independence are regarded by many as basic moral and political values (e.g. Hayek 1960; Berlin 1969). They are related to the notions of personal autonomy and liberty that philosophers and social psychologists have linked to well-being and demands of social justice.[2] For instance, Rawls (1971) uses

[2]See Christman (2011) and Anderson (2013) for reviews on different conceptions of autonomy, and the role of self-reliance and independence on the constitution of personal autonomy.

the conception of the autonomous person to formulate and justify his principles of justice; and Deci and Ryan (1985) contend that autonomy is essential for well-being. On the other hand, power as an end in itself (or what has been called the *love of domination*) is seen by many as an obstacle to freedom. For instance, Machiavelli (2003) saw the ambition of the powerful directed against the populace as the gravest and least easily neutralized danger to free governments (see Skinner 1992).

The first objective of this chapter is, therefore, to measure the weight of each of these rationales in the potential intrinsic value of decision rights.

In addition, the expression of an intrinsic value of decision rights and its underlying reasons may be linked to identifiable social conditions, contextual factors, and personal characteristics. In this respect, empirical evidence and casual observation suggest that the *cultural background* may have a strong influence when it comes to *core values* such as autonomy and power. There is a long list of cross cultural experimental studies that document significant differences in preferences and behavior across societies (e.g. Roth, Prasnikar, Okuno-Fujiwara and Zamir 1991; Henrich, Boyd, Bowles, Camerer, Fehr and Gintis 2004; Gachter, Herrmann and Thoni 2010). Preferences seem to be susceptible to framing, elicitation, anchoring and identity primes that are found in the social institutions and interaction patterns of a society (Fehr and Hoff 2011). As a consequence, institutions can be designed such that values and attitudes such as autonomy and power are favored or not. Priming one or the other, depending upon the context and society in question, may lead to different social and economic outcomes.

The second objective of this chapter is, therefore, to test if the intrinsic value of decision rights and its rationales are shared by different cultural backgrounds.

In order to achieve these two objectives, we extend the experimental design of Bartling et al. (2014) and implement, both in France and in Japan, a series of treatments that allow us to disentangle how much of the intrinsic value of decision rights stems from (i) the aversion to be affected by the decision made by someone else (*in-*

dependence), (ii) the desire to decide on behalf of someone else (*power*), or (iii) other motives such as the will to implement a decision that is the result of one's rational deliberation (*self-reliance*).

The choice of these two countries allows us to relate with the literature on the differences in core values between the *Eastern* and *Western* cultural backgrounds. For example, several studies suggest that East Asian cultures tend to place a greater emphasis on collectivism and cooperation than Western cultures that are based on more individualistic values (e.g. Markus and Kitayama 1991; Parks and Vu 1994; Wong and Hong 2005; LeBoeuf, Shafir and Bayuk 2010). Our experiment can test if the French (linked to Western) and Japanese (linked to Eastern) cultural backgrounds translate into different tastes for holding control, in particular in terms of the rationales (different core values) behind the potential intrinsic value of decision rights.

All treatments maintain the experimental design of Bartling et al. (2014), which consists of two parts. In the first part of the experiment, a principal has the choice between keeping a decision right or delegating it to an agent. This part is constructed in order to elicit the principal's point of indifference between keeping and delegating the decision right. In the second part, the monetary values of keeping and delegating control are elicited with a pair of lotteries that are constructed based on the choices made by the principal in the first part.

The treatments vary with respect to the nature of the agent. In the first treatment, which is a replication of the experiment by Bartling et al. (2014), the agent is a human subject. In the second treatment, a (ro)bot plays the role of the agent. Then, a principal who reveals an intrinsic value to hold control is not motivated by neither a preference for independence *nor* a desire for power. The principal cannot be neither affected by the decisions or exercise power over *another person*. This means that in the second treatment the potential intrinsic value of decision rights derives from other motives such as a preference for self-reliance. In the third treatment, a bot makes the decisions on behalf of a passive human agent. Here, an intrinsic value for holding control can be motivated by a preference for power or self-reliance but not

by a preference for independence from the interference of another person. This is the case since the principal cannot be affected by a decision made by any human agent, but the principal can hold control over the decisions that affect the passive human agent. The net effects of these treatments give us precise estimates to test the role of the different motives in the potential intrinsic value of decision rights.

We find that principals in both countries assign an economically and statistically significant intrinsic value to hold control over decisions. This suggests that the intrinsic value of decision rights may be a preference shared by individuals with Western and Eastern cultural backgrounds. It also brings support, as a cross-cultural replication, to the main finding of Bartling et al. (2014). Second, we find that on average, and taking all rationales into account, French principals attach a higher intrinsic value to decision rights than Japanese principals.

In terms of the motives for this preference, our main finding is that a preference for self-reliance is the only (positive) rationale behind the intrinsic value of decision rights in both France and Japan. This effect is economically and statistically significant, and of similar size in both countries. This means that, somewhat surprisingly, independence and power are not motivations behind the intrinsic value of being in control in our setting.

Clearly, holding control will not be *always* intrinsically valued. For instance, decisions may have an "unpleasant" component attached to them when they run counter to the interests of another person (Bartling and Fischbacher 2012). Similarly, our findings are not indications that power and independence are not motives for a preference for holding control and sources of value in principal-agent interactions. For instance, power may be valuable by the amenities it brings, such as status and recognition. What our results suggest is that without these additional instrumental sources of value, power and independence are less valued than a preference for self-reliance as motivations for the intrinsic value of decision rights.

The remaining of the chapter is organized as follows. We proceed with a brief relation to the literature. Section 5.3 is devoted to the experimental design. In Section

5.4 we describe the measurement of the rationales and our theoretical predictions. We report our findings in Section 5.5. We discuss some potential worries and limitations of the experimental design and our treatments in Section 5.6. Section 5.7 concludes.

5.2 Relation to the Literature

Our findings can be contrasted with the ones found by Neri and Rommeswinkel (2014), that look at the rationales behind the intrinsic value of decision rights with a different experimental design. On the one hand, our results match theirs in that a preference for power is not a strong determinant for the intrinsic value of decision rights.[3] On the other hand, our results contrast with theirs in terms of the relative importance of independence and self-reliance; in their experiment, independence is a stronger rationale for the intrinsic value of decision rights than self-reliance. The discrepancy may be due to the differences in the elicitation of the intrinsic value of decision rights and the definition and elicitation of the rationales behind this preference.

By replicating and refining the analysis started by Bartling et al. (2014), we contribute to the literature on incomplete contracts and the delegation of authority (e.g. Aghion and Tirole 1997). In particular, we add new evidence to the emerging literature on experimental economics on the value of decision rights, autonomy, and power (e.g. Falk and Kosfeld 2006; Fehr, Herz and Wilkening 2013; Owens, Grossman and Fackler 2014; Neri and Rommeswinkel 2014; Burdin, Halliday and Landini 2015). Bartling et al. (2014) show that underdelegation of decision rights from principals to agents, a result also found in previous experimentally controlled situations (e.g. Fehr et al. 2013; Owens et al. 2014), can be the result of an intrinsic value of

[3]Their experimental measure of power is somewhat different than ours since the player with the decision right is not choosing for the other player. In fact, the decision taken by the principal only randomly determines the agent's payoff. In our view, this should be interpreted either as (i) an illusion of power or (ii) a preference for limiting the freedom of the other.

holding control. Our results suggest that this underdelegation is mostly driven by a preference for self-reliance, rather than a preference for independence or power.

Likewise, our results contribute to the corporate finance and governance literatures (e.g. Aghion and Bolton 1992). Among the benefits of being in control, authors have acknowledged the "psychic value" that some shareholders attribute simply to being in control (Dyck and Zingales 2004, 540). Our analysis provides new empirical measures to test if these benefits are due to goals such as "the pursuit of power" (Hart and Moore 1995, 568). In addition, our results bring new insights over the rationales and social characteristics behind decisions such as self-employment and career choice. Non-pecuniary motives seem to be an important driver of decisions such as to enter self-employment (Hamilton 2000; Pugsley and Hurst 2011), and our results point towards the potential relevance of the ability to implement one's decisions against independence and power.

Given the cross-cultural dimension of our experiment, our analysis is also related to the research on the effect of the cultural background in shaping preferences and values. In particular, we contribute to the new strand of experimental research investigating the diversity of behavior and preferences across societies (e.g. Bohnet, Greig, Herrmann and Zeckhauser 2008; Henrich, Ensminger, McElreath, Barr, Barrett, Bolyanatz, Cardenas, Gurven, Gwako, Henrich, Lesorogol, Marlowe, Tracer and Ziker 2010; Gachter et al. 2010). Our results report new evidence on how the cultural background may influence core values that are often seen as fundamental for individual well-being and social welfare. As pointed by Fehr and Hoff (2011), having more information about the interaction between culture and preferences (values) may help policy makers to shape institutions that take into account the heterogeneity in preferences (values) that are found in different societies.

More specifically, our findings are relevant to the debate on the relationship between core values and the Eastern and Western cultural backgrounds. In previous studies, Asian-American subjects have been shown to have a higher tendency to cooperate when their Asian identity is made salient (LeBoeuf et al. 2010), and Asian-

American children, contrary to Anglo-American ones, have been shown to perform better - have higher intrinsic motivation - when a simple task is chosen by their mothers than by themselves (Iyengar and Lepper 1999). Following some of these and other results, some authors have challenged the held view that autonomy is an essential component of well-being for human beings worldwide (e.g. Iyengar and DeVoe 2003).[4] Even though our analysis cannot answer this deep question, our findings support the view that it is important to distinguish between autonomy as independence from an unknown human influence and autonomy as a sense of self-reliance based on the will to implement one's decisions (see Ryan and Deci 2006).

5.3 Experimental Design

In what follows, we summarize the experimental design of Bartling et al. (2014), explain the new treatments, and present the strategy used for the elicitation of preferences.[5] We end this section with details about the procedures followed in the experiment, notably on the ones used to ensure the robustness of our cross-cultural comparison.

5.3.1 Part 1: The Delegation Game

In the first part of the experiment, subjects play several 2-player one-shot delegation games. A *principal* (she) has the possibility to keep or delegate a decision right to an *agent* (him), which grants the right to implement a risky project. The outcome of the project determines the principal's and the agent's payoffs. There are two risky projects, $\mathcal{A}$ and $\mathcal{P}$, with the payoffs summarized in Table 5.1 where the subscripts refer to the projects, and 0 stands for the failure of the risky projects. The projects are risky since their success depends on the effort choice (probability of success)

[4] See Ryan and Deci (2006) for a response to these critics and a defense of autonomy as a *universal value*.

[5] We refer the reader to Bartling et al. (2014) for details and the theoretical foundation of the experimental design.

Table 5.1 – Projects' Payoffs

Principal (P)	Agent (A)
$P_{\mathcal{P}} \geq P_{\mathcal{A}} > P_0$	$A_{\mathcal{A}} \geq A_{\mathcal{P}} > A_0$

chosen by the player that keeps/gets the decision right. Before explaining how the delegation of decision rights from the principal to the agent takes place, we describe how the principal and the agent choose the project and its success probability.

The principal and the agent choose simultaneously a project, $\mathcal{A}$ or $\mathcal{P}$, and an *intended effort level*, denoted by E and e respectively, that they would like to implement in case they have the decision right. Effort can be chosen from the set $[0, 100]$ and corresponds to the probability (in percent) that the project will be successful. Effort is equally costly to the principal and the agent. The cost of effort is given by $C(E) = kE^2$ and $C(e) = ke^2$ respectively, where $k \in \{0.01, 0.02\}$ is a cost parameter.[6] Both subjects make their two choices simultaneously without knowing who is going to have the decision right, i.e., without knowing whose decision (of project and effort) will be implemented. Their choices are binding, and the ones made by the subject who ultimately holds the decision right determine the project to be implemented and its outcomes. Only this player pays the cost of effort.

In addition, and essential for the design, the principal also indicates a *minimum effort requirement* $\underline{e} \in [1, 100]$ that the agent needs to choose in order for her to be willing to delegate the decision right. That is, delegation takes place if and only if the agent's intended effort level is at least as high as the principal's minimum effort requirement, i.e., if and only if $\underline{e} \leq e$. The principal chooses this minimum requirement without knowing the e chosen by the agent; similarly, the agent chooses his intended effort level without knowing the $\underline{e}$ chosen by the principal. Then, *the minimum effort requirement should represent the principal's point of indifference between keeping the decision right and delegating it to the agent*.

[6] The cost parameter k varies across rounds, but it is always common knowledge and identical for both players.

Table 5.2 – Parameters of the Games

	Project Successful				**Project**			
	Project $\mathcal{P}$		**Project $\mathcal{A}$**		**Unsuccessful**			
	$P_{\mathcal{P}}$	$A_{\mathcal{P}}$	$P_{\mathcal{A}}$	$A_{\mathcal{A}}$	P_0	A_0	$C(E)$	$C(e)$
Game 1	220	190	190	220	100	100	$0.01E^2$	$0.01e^2$
Game 2	280	235	235	280	100	100	$0.01E^2$	$0.01e^2$
Game 3	180	140	140	180	100	100	$0.01E^2$	$0.01e^2$
Game 4	220	160	160	220	100	100	$0.01E^2$	$0.01e^2$
Game 5	260	260	260	260	100	100	$0.01E^2$	$0.01e^2$
Game 6	440	380	380	440	200	200	$0.02E^2$	$0.02e^2$
Game 7	560	470	470	560	200	200	$0.02E^2$	$0.02e^2$
Game 8	360	280	280	360	200	200	$0.02E^2$	$0.02e^2$
Game 9	440	320	320	440	200	200	$0.02E^2$	$0.02e^2$
Game 10	520	520	520	520	200	200	$0.02E^2$	$0.02e^2$

In total, subjects play ten different delegation games (ten rounds) with stranger matching. The participants remain in the role of principal or agent throughout the experiment, and receive no feedback about the outcomes in a given round until the end of the experiment. The rounds differ only with respect to the projects' payoff and costs of effort. These differences allow us to test for situational determinants of a potential intrinsic value of decision rights such as the *stake size* and the *conflict of interest* between the principal and the agent. Table 5.2 summarizes the main parameters of the ten games.[7]

Subjects are paid at the end of the experiment according to the outcome of one randomly chosen round. Payments, that depend upon the success of the project, are based on a random number $r \in [1, 100]$ and the effort level chosen by the participant with the decision right. If the principal has the decision right, the project is successful if $r \leq E$; if instead the agent has the decision right, the project is successful if $r \leq e$. Otherwise (if neither of these cases is realized), the project fails.

[7] The order of the games/rounds was randomized across sessions.

5.3.2 Part 2: The Lottery Task

In the second part of the experiment, subjects perform an individual decision task. Each subject states their certainty equivalent for each of 20 different lotteries, i.e., the smallest certain payoff they are willing to accept instead of each one of these lotteries. Each lottery determines probabilistically the subject's own payoff and the payoff of another randomly paired participant.

Consider the subjects who played the role of the principal in Part 1 of the experiment. These are the subjects of interest, since Part 2 is an individual task, and they are the ones for which we measure their potential intrinsic value of holding control. For each of these subjects, and *without them being aware of that*, the lotteries are constructed based on their own choices in the delegation game.[8]

In each round of the delegation game, their choice of $\mathcal{P}$, E, and $\underline{e}$ determine a pair of lotteries: (i) A principal's intended effort (with the corresponding effort cost) and the chosen project fully determine a *control lottery*; and (ii) her minimum effort requirement fully determines a *delegation lottery*.[9]

For each lottery, the subjects' certainty equivalent is elicited in an incentive compatible manner *à la* Becker, DeGroot and Marschak (1964). The lotteries consist of a low ($\underline{P}$) and a high ($\overline{P}$) payoff for the principal and a low ($\underline{p}$) and a high ($\overline{p}$) payoff for another randomly paired participant.[10] The lotteries are presented in an individ-

[8]Withholding this information is essential to separate the elicitation of Part 2 from the decisions in Part 1. We consider that withholding this information is acceptable (contrary to plain deception), mainly because it should have no bearing neither on the subjects' experience nor outcomes. Since the order of the lotteries (and rounds in Part 1) is random and their connection is not self-evident, it seems also very unlikely that any subject has become aware of this fact. A potential issue with not disclosing the details of all parts since the beginning of the experiment is that subjects in Part 1 might think that their decisions will matter later. Though this is a potential concern, it should not affect our treatment comparisons. Participants who were in the role of agents in Part 1 perform the same task as principals in Part 2 but the lotteries that are presented to them are based on the decisions taken by the principals with whom they are associated.

[9]Condition (ii) is grounded on the assumption that the principal perceives delegation to lead to the choice of project $\mathcal{A}$ (i.e. the principal anticipates that the agent will/may choose the project that gives him the higher payoff). We discuss this assumption in Section 5.6.

[10]This differs from a typical experimental certainty equivalent elicitation task since lotteries and certainty equivalents involve not only payoffs for the decision maker but for two parties. This is done to ensure comparability with the lotteries in Part 1.

ually randomized order, and principals are asked to state the smallest certain payoff - the certainty equivalent (ce) - that they are willing to accept instead of the lottery.

At the end of the experiment, two of these 20 lotteries are randomly chosen to be relevant for payment. Each of these two lotteries is either played or not depending upon the ce that the principal has stated for that lottery and a randomly generated number $r \sim U[\underline{P}, \overline{P}]$. If $r \geq ce$ the principal receives r for sure and the randomly paired participant receives a certain payment equivalent to those of the projects in Part 1 in case of failure.[11] Otherwise (if $r < ce$), the lottery is played. Feedback is given to participants only at the end of the experiment.

5.3.3 Example of Parts 1 and 2

To summarize, take the following example. Suppose that instead of ten games, there was only Game 1 in Part 1 (see Table 5.2), and assume that a principal chooses project $\mathcal{P}$, $E = 50$ (with a corresponding cost $C(E) = 25$), and $\underline{e} = 40$ (with an associated effort cost $C(e) = 16$ for the agent). Assume that her randomly matched agent chooses project $\mathcal{A}$ and $e = 30$ (with a corresponding cost $C(e) = 9$). According to these choices, the principal keeps the decision right; but since feedback is not yet given to the participants, Part 2 starts without the subjects being aware of the outcomes of Part 1. Two lotteries - with payoffs for the principal and another randomly paired participant - are determined from the principal's decisions taken in Part 1 without them being informed of it:

- A control lottery such that $\overline{P} = 220 - 25 = 195$ and $\overline{p} = 190$ with 50% probability (high payoff), and $\underline{P} = 100 - 25 = 75$ and $\underline{p} = 100$ with 50% probability (low payoff).

[11] More precisely, the randomly paired participant receives 100 points in the lotteries derived from the choices in games 1 to 5 and 200 in the ones derived from the choices in games 6 to 10.

- A delegation lottery such that $\overline{P} = 190$ and $\overline{p} = 220 - 16 = 204$ with 40% probability (high payoff), and $\underline{P} = 100$ and $\underline{p} = 100 - 16 = 84$ with 60% probability (low payoff).

The principal then states her certainty equivalents (ce) for these two lotteries. Assume that the principal states a certainty equivalent for the control lottery of 140 and a certainty equivalent for the delegation lottery of 160. Payments are then calculated according to three randomly generated numbers, say r_1 (for Part 1), and r_2 and r_3 (for Part 2).[12] If $r_1 \leq E$ the principal receives $220 - 25 = 195$ points and the randomly matched agent receives 190 for Part 1. Otherwise (if $r_1 > E$), the principal and the agent receive 75 and 100 points for Part 1 respectively. For Part 2, if $r_2 \geq 140$ for the control lottery then the principal receives r_2 for sure and the randomly paired participant receives 100 points. Otherwise the control lottery is played and payments are determined according to its result. For the delegation lottery, if $r_3 \geq 160$ then the principal receives r_3 for sure and the randomly paired participant receives 100 points. Otherwise the delegation lottery is played and payments are determined according to its result.[13] In addition, the principal receives a supplementary payment for her role as a "randomly paired participant" to another subject in Part 2. Feedback is finally given to all participants.

5.3.4 Treatments

To examine the rationales behind a potential intrinsic value of decision rights, we implemented three treatments (T1, T2, and T3) that vary the nature of the agent. In T1, the agent is a human subject. In T2 a bot (computer program) plays the role of the agent.[14] And in T3 a bot makes decisions on behalf of a passive human agent. Figure 5.1 summarizes the organization of these treatments.

[12] The same holds for the full experiment, except that the game of Part 1 and the lotteries of Part 2 are chosen independently for payment.

[13] Note that the results of the control and delegation lotteries, if they are played, will be based on randomly generated numbers. We have omitted them to simplify the exposition.

[14] In the experiment we use the term "bot" (instead of computer program) to harmonize the instructions of the experiment with relation to T1.

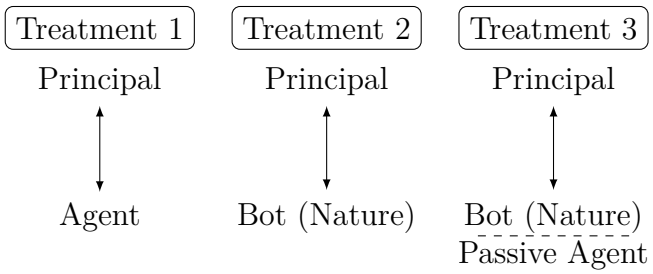


Figure 5.1 – Treatments

In Part 1 of T2 and T3, the bot (i) always chooses the project alternative that provides it (or the passive human agent) the larger payoff (i.e., project $\mathcal{A}$), and it (ii) uniformly randomly determines its effort from $[0, 100]$. The principal (and the passive human agent) are informed of this decision making model of the bot. Condition (i) is consistent with the theoretical (self-interested) prediction for the agents' choice in Part 1. Condition (ii) is an attempt to mimic *nature*, in the sense of a chance device. This is also the interpretation given to a random device by Bohnet et al. (2008), that use a similar strategy to disentangle different rationales behind subjects' willingness to take risk. In Part 2, the bot randomly determines the certainty equivalents. This has no bearing to our design since Part 2 is an individual decision task in which agents (and the bot) are only known to the principal in the position of randomly matched "participants" that are affected by her decisions.

In T3, if the bot receives the decision right, it makes the decisions on behalf of the passive human agent. To put it another way, the passive human agent makes no relevant decision themselves.[15] This means that the passive human agent's payoff depends on the decisions taken by the principal and the bot, while the principal's payoff depends on their own decisions and the ones taken by the bot. In Section 5.6 we discuss the potential implications of social and risk preferences, ambiguity aversion, beliefs about the agent's behavior, and differences in beliefs regarding the agent's and bot's behavior.

[15] In order to entertain these participants, we let them perform the same decisions as agents in T1 although they are aware that these decisions will not be taken into account for payoffs.

5.3.5 Additional Experimental Measures

After Part 2 of the experiment, we ask subjects to complete a series of short tasks. As in Bartling et al. (2014), we elicit participants' loss aversion and illusion of control. In our experiment, we also elicit participants' cognitive ability. These measures allow us to control for alternative explanations for a potential intrinsic value of decision rights.

We follow Bartling et al. (2014) and elicit the subjects' degree of loss aversion using a lottery task (a design taken from Fehr and Goette 2007). Subjects accept or reject a series of lotteries involving possible losses of different sizes X. For example, in France participants either accepted or rejected lotteries with a 50% probability of winning 5€ and 50% probability of losing X€, with X going from 1 up to 6€ (i.e., $X \in \{1, ..., 6\}$).[16] The amount X at which a participant starts rejecting the lotteries is an indicator of his or her loss aversion.[17] For instance, a participant who rejects all lotteries with a potential loss of $X > 2$ is classified as more loss averse than a participant who only rejects lotteries with a potential loss of $X > 5$.

To elicit subjects' illusion of control, we adopt a modified version of the incentive compatible elicitation method used by Charness and Gneezy (2010). We measure illusion of control as the principal's willingness to pay for the right to personally stop the roll of two ten-sided electronic dice (that determine the random outcomes in Part 1 and 2 of the experiment).[18] The key insight is that if principals are subject to an illusion of control, they should value stopping the rolling of the dice because this increases their *perceived* personal involvement in determining the final outcomes. If they opt not to personally stop the rolling dice, the dice stop automatically.

[16] If they accepted, the lottery was played, otherwise they received 0€. Once all decisions are taken, one of the six lotteries is randomly selected for actual payment and, in case of acceptance, a computerized random draw determines its outcome.

[17] Remark that we might be unable to ascertain a participant's loss aversion if he or she has not a unique switching point. In our sample, there were 4 "non-consistent" subjects.

[18] We modify the task by substituting the two physical dice by two electronic dice that appear on the participants' screen. Principals are asked if they are willing to pay to personally stop the rolling dice. They are informed that the numbers change too quickly for them to be able to choose which numbers to stop on. This avoids the time consuming activity (that involves the participants and the experimenter) of rolling the physical dice.

To assure the comparability between sites and for an additional control on the potential effects of bounded rationality, we elicit the subjects' cognitive ability through a *Raven's Progressive Matrices test* (RPM test). Recent experimental studies show that the scores of RPM tests are correlated with subjects' behavior in strategic games (see e.g. Burks, Carpenter, Goette and Rustichini, 2009, Carpenter, Graham and Wolf, 2013, Gill and Prowse, 2016, Hanaki, Jacquemet, Luchini and Zylbersztejn, 2016a,b). This test is widely used worldwide, and is especially suited for cross-cultural studies since it is independent of language, reading, and writing skills. The test consists of choosing among a given number of patterns the one that best fits the "blank space" of a visual geometric design. The number and difficulty of the visual geometric designs vary from one version of the test to the other. In our experiment, subjects were asked to choose among 8 patterns (8 options) and there were 16 different visual geometric designs of distinct difficulties, taken from the advanced version of the RPM test (Raven, 1998), to be answered in 10 minutes. Our measure of cognitive ability is then the score of this task computed as the number of correct answers.

5.3.6 Procedures

All subjects were students recruited at universities where we have conducted our experiments. The experiment was computerized using z-Tree (Fischbacher 2007). Payments were made for one randomly drawn round of the delegation game (Part 1), and for four randomly drawn lotteries in Part 2 (two of them in the role of the "random other participant"). Subjects received an extra endowment for the loss aversion and illusion of control tasks and were paid according to their results on these tasks. The following exchange rate applied: 100 points = 2.5€ or 300 Yen. Subjects earned a 5 € or 600 Yen show-up fee and received on average an additional 29.6€ in France and 3450 Yens in Japan in experimental sessions that lasted on average 2.5 hours.

Participants were provided with paper-based instructions for all parts of the experiment. Instructions for Part 2 of the experiment were handed out only after Part 1

was finished. Participants knew that the experimental session would consist of several parts, but they did not know the precise content of the future parts before the respective instructions were provided. To ensure that subjects understood the experimental design and the impact of their decisions on their earnings, they had to answer a series of control questions after reading the instructions of Part 1 and after reading the ones of Part 2. They were confronted with the different choices they would have to make during the experiment, and their answers were corrected and shown to them for revision.

To ensure the equivalence of experimental procedures across countries, we followed (for the most part) the methodology first used and described in Roth et al. (1991). We try to control for (i) subject-pool, (ii) language, (iii) currency, and (iv) experimenter effects. To control for *subject-pool effects*, we recruited only university students in both locations. Our subject pool is then mostly homogeneous in terms of length of educational background. In addition, we conducted a cognitive test in the end of the experiment (as explained in the previous section) to control for potential effects of differences in cognitive abilities. Finally, university students of non-western countries may be among the most "westernized" individuals of these countries. Our results can then be seen as a lower bound in terms of the effects arising from cultural differences. To minimize *language effects*, instructions in English were translated into local language by a French or Japanese native speaker, and back translated to English by another person. Translators were careful to write the instructions in neutral language, and the authors ensured compatibility with the German and English instructions of Bartling et al. (2014). In terms of *currency effects*, payoffs were expressed in "points" and the comparability of earnings was ensured by taking an average between standards of living, local hourly payments, and show-up fee practices of the laboratories.[19] To minimize *experimenter effects*, all experimenters were native speakers and were present in the first session of each treatment that were ran

[19]With the exception of the loss-aversion task where outcomes were expressed in the local currency (as in Bartling et al. 2014). The conversion rate was conserved.

in France. The experimental sessions in Japan were conducted by the second author, and the experimental sessions in France were conducted by the third. Further, most of the instructions were read individually, minimizing the subjects' interaction with the experimenter.

5.4 Motives and Culture: Measurement and Predictions

Regardless of the unobserved social and risk preferences of the principal, the experimental design ensures that it is optimal for the principal to choose a minimum effort requirement $\underline{e}$ such that she is indifferent between keeping and delegating the decision right at $\underline{e}$. As discussed below (Section 5.6), the optimal choice of $\underline{e}$ should be independent from the principal's beliefs about the agent's/bot's effort choice. Then, the intrinsic value of decision rights (IV hereafter) is measured by comparing the certainty equivalents (monetary values) of the control and delegation lotteries. The principal's utility of keeping control consists of the monetary value of the control lottery [ce(CL)] *plus* the potential IV. The principal's utility of delegating at $\underline{e}$ consists of the monetary value of the delegation lottery [ce(DL)]. As shown by Bartling et al. (2014), it is then possible to quantify the potential IV as the certain amount of points (money) that a principal demands as a compensation for the delegation of the decision right:

$$\text{IV} = ce(DL) - ce(CL)$$

We distinguish three motives that can underlay a potential intrinsic value of holding control. The first motive is the aversion to be affected by the decision made by someone else. This can be seen as a preference for *independence* (R1) and is closely connected to the concept of negative freedom. In one of its most famous formulations, Berlin (1969) defines negative freedom as the freedom from constraints that are imposed by others (as opposed to constraints such as biological impediments).

For Berlin (1969, 8), negative freedom consists in "not being prevented from choosing as I do by other men." Similarly, R1 can be interpreted as the will not to be subordinated or to enjoy *freedom from the interference* of others.

The second motive is the desire to be able to decide on behalf of someone else. This can be interpreted as a notion of *power* (R2), when power is defined as authority exercised over others. For instance, Simon (1951, 294) states that "B exercises *authority* over W if W permits B to select x [a part of his/her behavior]".[20] This rationale can be also interpreted as what has been called the *love of domination*, that Mercy Otis Warren claimed to *have prevailed among all nations and perhaps in proportion to the degrees of civilization*.

Finally, we highlight a preference for *self-reliance* (R3) as a candidate rationale for the intrinsic value of holding control. This is an important component of personal autonomy, and can be seen as the capacity of "being able to realize one's intentions and goals" (see Anderson 2013, 6). According to Raz (1986, 369), "[t]he ruling idea behind the ideal of personal autonomy is that people should make their own lives". If this rationale is behind the IV in our experiment, then keeping the decision right is valued for the desire to implement one's decision. In this sense, we can distinguish between the traditional negative view of freedom/independence from others (as in R1) and a concept of "inner freedom" related to a sense of self-reliance and control (as in R3). While self-reliance has a strong internal character that our experiment is not able to grasp, the design allows us to test if a preference for personal autonomy in this sense may be a reasonable determinant of the IV.

To measure the strength of these different motives, we compare the net effects between the three treatments. R1, R2, and R3 are potential rationales for the IV in T1, R3 is the sole potential rationale behind the IV in T2, and R2 and R3 are potential rationales for the IV in T3. Taking the measured IV of our three treatments, we can

[20] Similarly, according to Dahl (1957, 202-03) "A has power over B to the extent that he can get B to do something that B would not otherwise do". This is the most common notion in modern political theory, instead of power as the ability to do or act (see Patton 1989). See Bowles and Gintis (1988) for the relevance of the treatment of power in economic theory.

construct precise estimates of the weight of each motive in the potential intrinsic value of holding control:

$$\begin{aligned}
\textbf{R1} &= IV_{T1} - IV_{T3} = (R1 + R2 + R3) - (R2 + R3). \\
\textbf{R2} &= IV_{T3} - IV_{T2} = (R2 + R3) - (R3). \\
\textbf{R3} &= IV_{T2} = (R3).
\end{aligned}$$

We predict that all these rationales have a significant and positive impact on the intrinsic value of holding control. We expect independence from interference (R1) to have a positive impact given the long philosophical tradition on the value of negative freedom. For instance, this notion of freedom is endorsed as essential by most liberals, such as Hayek (1960), Nozick (1974), and Buchanan (1986). Power (R2) is expected to be a rationale behind the IV since it is seen by many as a powerful motivation for human behavior. For instance, according to McClelland's (1975) motivation theory, power is a dominant human need. Finally, self-reliance (R3) is expected to be a motive for the IV since it is often associated with increased well-being and a sense of worth. For instance, Deci and Ryan (1985) hold that self-determination/autonomy is essential for well-being, and the freedom of choice literature resonates over the idea that the process of choice itself has an intrinsic worth and that the opportunity to choose or act is essential for individuals to lead the lives they have reason to value (see e.g. Sen 1988). This can be resumed in the following hypotheses:

Hypothesis 1. $R1 > 0.$

Hypothesis 2. $R2 > 0.$

Hypothesis 3. $R3 > 0.$

In terms of cultural differences, we predict that principals in France attach a higher intrinsic value to be in control than principals in Japan. Our hypothesis is

that this is driven by a higher value of independence, i.e., that French give more relevance than Japanese to not be affected by someone else's decision. This seems to be consistent with the evidence that Western individuals are prone to endorse more individualistic values than Eastern individuals (e.g. Markus and Kitayama 1991; Parks and Vu 1994; Wong and Hong 2005; LeBoeuf et al. 2010). It is also supported by the World Values Survey (WVS) 2015 data that suggests that French give considerably more weight to *self-expression* values than Japanese.[21] This hypothesis can be written as follows:

Hypothesis 4. $R1(France) > R1(Japan)$.

Cultural differences in terms of preferences for power and self-reliance are unclear. First, to the best of our knowledge there is few theoretical analysis on the cultural/societal determinants of these preferences. Second, the empirical evidence is scarce and for the moment inconclusive. Although some authors argue that more individualistic societies may foster a higher preference for control over others (e.g. Lee 1997), for now there is no accumulated evidence on this front. With regard to self-reliance, its "inner" nature and the non-settled dispute about the worldwide value of personal autonomy favor the absence of an *a-priori* position. Accordingly, we make no prediction about the different weights of these motives between France and Japan.

5.5 Results

This study involved 521 subjects, from which 319 were in the position of principal. Since we are interested in potential differences due to the cultural background, we drop from the analysis 45 principals that are neither of French or Japanese Nationality nor born in France or in Japan.[22] From the remaining 274 principals (observa-

[21] See e.g. Inglehart and Baker (2000) for theoretical background on the WVS data and the "self-expression versus survival" measure.

[22] See Appendix 5.E for other restrictions based on "stronger" and "weaker" definitions of link to the country or cultural background. Our results are robust to the different definitions.

tions), 142 participated in France [94 in Rennes (28 in T1, 38 in T2, 28 in T3) and 48 participated in Nice (15 in T1, 20 in T2, 13 in T3)] and 132 participated in Japan [65 in Osaka (23 in T1, 18 in T2, 24 in T3) and 67 in Tsukuba (20 in T1, 25 in T2, 22 in T3)].[23]

In the final part of the experiment, we collected some demographic, lifestyle, and values measures that can bring some insights into the similarities and differences of the samples of the two countries. In terms of demographics, the median age in France was 20 years old and 21 in Japan. There were 67% female subjects in France and 37% in Japan. In terms of fields of study, the French sample is more homogeneous than the Japanese one. In particular, while 57% of French subjects were students in economics and management only 27% of Japanese subjects were in the most representative field in the Japanese sample (Humanities). In a self-assessed social class scale from 1 (Upper class) to 5 (Lower class), the average was the same in both countries and equal to 2.7. In terms of cultural dimensions, 56% in France and 81% in Japan reported not to belong to any religion or religion denomination.[24] The most representative religions were "Buddhist" (12% of total subjects) and "other" (5%) in Japan, and "Roman Catholic" (22%) and "Muslim" (15%) in France. On average, subjects in France self-positioned themselves in a political scale more to the left than subjects in Japan. On a scale of 1 (extreme left) up to 10 (extreme right), the mean position was 4.8 in France and 5.6 in Japan. Furthermore, 70% of subjects in France self-positioned themselves from 5 to the left, while only 52% did so in Japan. In addition, while only 13% in France considered themselves to belong to a political score above 7, 30% of subjects reported so in Japan. In terms of (postmaterialist) liberty aspirations, on an index from 0 (low aspirations) to 5 (high aspirations), subjects in France have on average higher aspirations (mean of 2.2 with

[23] We ran a total of 33 sessions: 14 at Rennes (France), 8 at Nice (France), 5 at Osaka (Japan), and 6 at Tsukuba (Japan). The fewer observations in Nice are due to the nationality restriction.

[24] From the 44% remaining subjects in France, 8% declared that they did not wish to answer to this question.

only 11% of subjects with a score of 0) than subjects in Japan (mean of 1.5 with 30% of subjects with a score of 0).[25] Table 5.1 summarizes some of these characteristics.

These measures suggest that (i) the two samples – principals from France and Japan – are, with the exception of gender and the main field of study, similar in terms of non-cultural dimensions. They also suggest that (ii) the two samples may be representative of the French and Japanese cultural backgrounds. For instance, according to the World Values Survey data[26] of two representative samples of Japanese and French individuals, 58% of Japanese declared not to belong to any religion or religion denomination and 31% declared to be Buddhist, while in France 49% declared not to belong to any religion or religion denomination, 42% declared to be Catholic and 5% Muslim. These results, even though different than ours (which is expected since our sample consists only of university students), are generally aligned with ours. The results on political orientation are very similar to ours, with the mean position of 4.8 in France and 5.5 in Japan. In terms of liberty aspirations, Welzel and Inglehart (2005) report, based on the WVS of 1989-91 data, and as in our sample, higher liberty aspirations in France (around 2.6) than in Japan (around 2.15).

We also collected data on cognitive ability (Raven's score), loss aversion and illusion of control. In terms of cognitive ability, Japanese subjects seem to get a better score than French ones.[27] There is a certain homogeneity among French and Japanese subjects concerning degree of loss aversion: On average, the mean switch

[25] A subject is said to have high liberty aspirations if he or she mentions the following (national) goals - among others such as economic growth and maintaining order - as most important: "Seeing that people have more say about how things are done at their jobs and in their communities", "giving people more say in important government decisions", and "protecting freedom of speech". See Welzel and Inglehart (2005) for the theoretical justification of this index.

[26] The data reported is from the Wave 5 of the WVS, that collected data of 1096 Japanese individuals in 2005 and of 1001 French individuals in 2006.

[27] This difference is significant based on an OLS regression with other individual controls ($p < 0.001$). Previous studies have found that on average adult males score higher than adult females in this test across countries (see Lynn and Irwing 2004 for a meta-analysis). In our total sample, we find that being female decreases on average the Raven's score by 0.61 points based on an OLS regression without individual controls ($p < 0.10$), but this difference becomes insignificant when we control for other individual characteristics. Either controlling or not for individual characteristics, we find no significant variation within both countries (based on Wald tests). But while in Japan females have a 0.90 lower score than males, in France the opposite is true with females having on average a 0.69 higher score than males.

Table 5.1 – Characteristics of Subjects

	France	Japan
Age	20.1	21.1
Female (Fraction)	66.9%	37.1%
Social class (1 High, 5 Low)	2.7	2.7
Field of study:		
Economics & management	57.8%	18.2%
Law	14.1%	7.6%
Humanities	6.3%	27.3%
Math	4.2%	2.3%
Sciences	2.1%	18.2%
Other	15.5%	26.5%
Political scale (0 Ext. left, 10 Ext. right)	4.8	5.6
Not religious (Fraction)	56.3%	81.1%
Liberty aspirations (0 Low, 5 High)	2.2	1.5
Raven's score (Max. 16)	9.7	11.5
Loss aversion (1 High, 6 Low)	3	3.2
Illusion of control:		
Fraction W.T.P. = 0	64.9%	72%
Mean W.T.P. if W.T.P. > 0 (Max. 30 points)	10.6	13.2

Notes: Mean values for "Age", "Social class", "Political scale", "Liberty aspirations", "Raven's score", and "Loss aversion". W.T.P. refers to "willingness to pay" (see Section 5.3.5).

occurs for 3€, i.e., subjects reject the lottery with 50% probability of winning 5€ and 50% probability of losing 3€. As for illusion of control, 65% of French principals and 72% of Japanese ones chose not to pay for personally stopping the roll of two ten-sided electronic dice. Among those being willing to pay for rolling the dice, they accept to pay, on average, 10.6 of their 30 additional points in France and 13.2 in Japan.

Accordingly, in our analysis we control for gender, cognitive ability (Raven score), field of study, loss aversion, and illusion of control. This way we take into account the differences in gender and field of study compositions between locations and alternative explanations based on bounded rationality, loss aversion, and illusion of control.[28]

[28] We address the remaining alternative explanations discussed in Bartling et al. (2014) in Appendix 5.F.

In the rest of the section, we analyze the intrinsic value of decision rights as measured by the difference in certainty equivalents of the delegation and the control lotteries: IV $= ce(DL) - ce(CL)$. Whenever not mentioned differently, we use this as our IV measure. In Appendix 5.D we report some of our results using the percentage difference in certainty equivalents that normalizes the IV for the monetary stake of the lotteries.[29]

5.5.1 Within Country Differences

Since we have data from two locations in each country, we can bring some support to the claim that the two samples may be representative of the French and Japanese cultural backgrounds with behavioral data. In particular, we can check if the subjects' behavior is consistent *within each country*. If we observe no significant difference between two locations of the same country, we can be more confident that there is a certain degree of homogeneity within a country. This would bring some support to the claim that we are capturing the cultural background.

Note that this argument holds for *two locations of the same country* but not for *two locations of different countries*. We have no theoretical reason (in terms of definition of cultural background) to pool the data from two locations of different countries even if we find no significant difference for the subjects' behavior between two of these locations.

5.5.1.1 France

Figure 5.1 presents the IV measure for the three treatments in the two locations in France, based on the estimated coefficients and standard errors of the treatment-location dummy of an OLS regression controlling for individual characteristics re-

[29] We opt to report the former measure in the main text since we cannot use the latter to measure the weight of each rationale on the IV. In fact, using the percentage difference we cannot exclude that our measure of power is biased by the effect of social preferences. While our IV measure neutralizes the effect of social preferences for any treatment and comparisons between treatments, the percentage difference, given by $IV/CE = \frac{ce(DL)-ce(CL)}{(ce(DL)+ce(CL))/2}$, does not neutralize the effect when comparing a treatment with and one without social preferences, as with $T3$ and $T2$ for our measure of power.

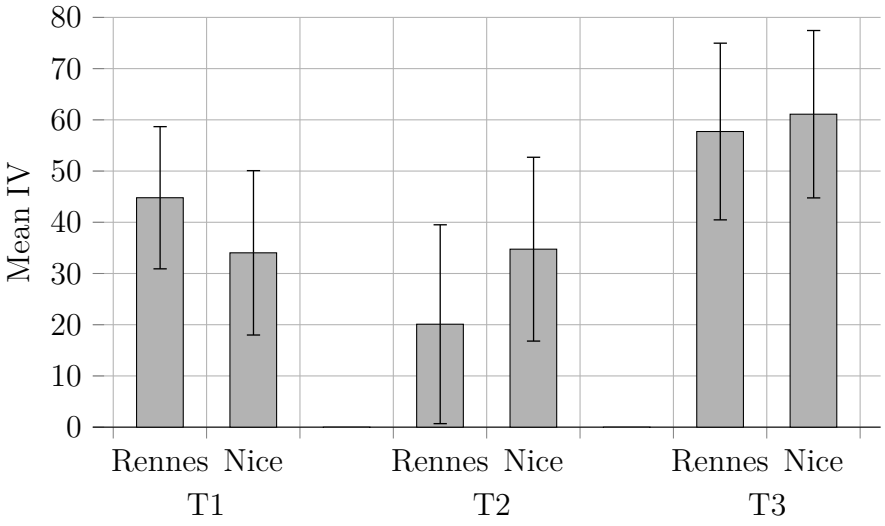


Figure 5.1 – Mean IV, sorted by French location and treatment. The bars display one standard error of the mean.

ported in Table 5.2.[30] As it can be seen from the figure, the values per treatment are similar for the two locations within France. Regression analyses reported in Table 5.2, either controlling or not for individual characteristics, do not reject the null hypothesis that the measured IVs in the two locations are the same for both T1 and T3. As for T2, we reject this hypothesis at 10% level when using the average IV and controlling for individual characteristics. However, either using nonparametric tests (see Table 5.B.1 in Appendix 5.B) or controlling for the stake size using the percentage difference (see Table 5.D.1 in Appendix 5.D) we do not reject the null hypothesis that the measured IVs in the two locations are the same for all treatments.

When looking at other behavioral data, in particular to the three decisions that principals made in Part 1, we find that the chosen minimum effort requirement $\underline{e}$ is very similar (specially in T1 and T2) and not statistically different for any treatment between the two locations (see Table 5.A.1 in Appendix 5.A), that effort choices are very similar and not statistically different for T2 and T3 but different and statistically different for T1 (see Table 5.A.2 in Appendix 5.A), and that principals chose the project that gave them the higher profit (project $\mathcal{P}$) in 87% of games in Rennes and 79% of games in Nice.

[30] All the results presented in this section are robust if we use two separate regressions for France and Japan.

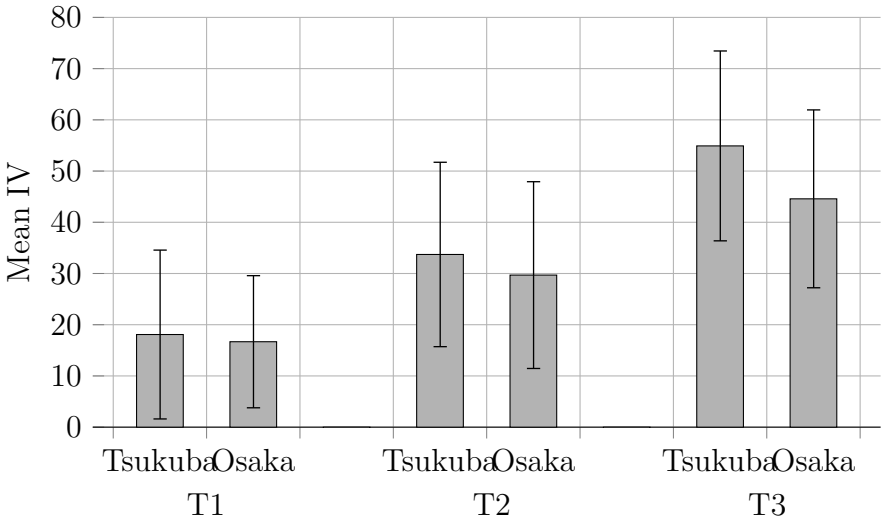


Figure 5.2 – Mean IV, sorted by Japanese location and treatment. The bars display one standard error of the mean.

5.5.1.2 Japan

As for Japan, Figure 5.2 presents the IV measure for the three treatments in the two locations based on the OLS regression controlling for individual characteristics reported in Table 5.2. As it can be seen from the figure, the values per treatment are very similar for the two locations within Japan. Regression analyses reported in Table 5.2, either controlling or not for individual characteristics, do not reject the null hypothesis that the measured IVs in the two locations are the same for all treatments. This result is robust either using nonparametric tests (see Table 5.B.1 in Appendix 5.B) or controlling for the stake size using the percentage difference (see Table 5.D.1 in Appendix 5.D).

In terms of the three decisions that principals made in Part 1, we find that the chosen minimum effort requirement $\underline{e}$ is very similar (specially in T1 and T2) and not statistically different for any treatment between the two locations (see Table 5.A.1 in Appendix 5.A), that effort choices are not statistically different for T1 and T2 but different and statistically different for T3 (see Table 5.A.2 in Appendix 5.A), and that principals chose the project that gave them the higher profit (project $\mathcal{P}$) in 89% of games in Tsukuba and 87% of games in Osaka.

Table 5.2 – Within Country Differences, *IV*

	Treatment 1		Treatment 2		Treatment 3	
Rennes	46.454***	44.792***	32.279***	20.098	41.221***	57.721***
	(5.488)	(13.880)	(5.648)	(19.413)	(5.395)	(17.248)
Nice	34.460***	34.036**	45.125***	34.749*	45.769***	61.101***
	(7.376)	(16.047)	(5.603)	(17.946)	(9.205)	(16.341)
Tsukuba	20.660***	18.086	40.440***	33.716*	33.018***	54.912***
	(7.686)	(16.482)	(4.222)	(18.001)	(5.318)	(18.535)
Osaka	17.430***	16.679	39.311***	29.689	29.762***	44.577**
	(6.058)	(12.899)	(4.935)	(18.233)	(6.353)	(17.364)
Female		−1.128		1.521		1.142
		(7.142)		(6.902)		(6.807)
Raven's Score		−0.082		0.068		−0.772
		(0.973)		(1.129)		(1.384)
Economics & Management		−3.331		4.340		−3.566
		(6.902)		(6.641)		(7.252)
Loss Aversion		0.757		1.826		−3.189*
		(1.680)		(2.499)		(1.820)
Illusion of control		0.313		0.389		0.219
		(0.424)		(0.440)		(0.405)
R^2	0.174	0.169	0.250	0.254	0.260	0.271
N	860	850	1010	990	870	860
H_0: Rennes = Nice	0.196	0.236	0.110	0.074	0.671	0.759
H_0: Tsukuba = Osaka	0.742	0.900	0.862	0.556	0.695	0.205

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regressions with clustered standard errors per subject.

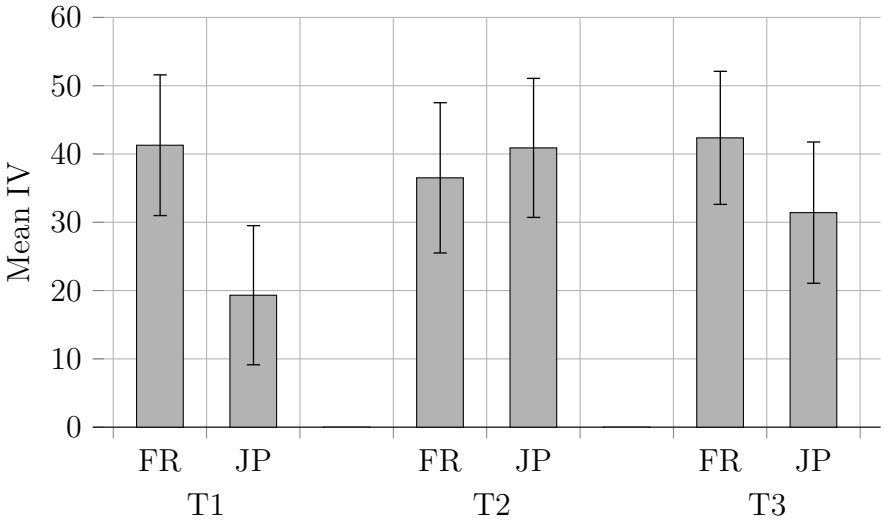


Figure 5.3 – Mean IV, sorted by country and treatment. The bars display one standard error of the mean.

We take these results as a suggestion that the subject-pools within each country can come from the same distribution. The lower degree of homogeneity of behavior within France when compared with the higher homogeneity within Japan may result from the more multicultural and varied setting of France and the higher social homogeneity of Japan. Taking the above values measures and these results into account, in what follows we perform the analysis with pooled data by country.

5.5.2 The Intrinsic Value of Decision Rights

We start our main analysis by comparing the IV between the two countries. Figure 5.3 presents the values of the IV measure for the three treatments in each country. These values are based on an OLS regression with individual controls reported in Table 5.3.

Result 1. *Decision rights have on average a positive intrinsic value for both French and Japanese principals.*

As can be seen in Figure 5.3, the certainty equivalents of the delegation lotteries are on average 19 to 42 points higher than those of the control lotteries depending on the treatment and country. This amounts to 13% to 25% in terms of percentage

Table 5.3 – The Intrinsic Value of Decision Rights, *IV*

T1 France	42.270***	41.284***
	(4.469)	(10.304)
T1 Japan	18.933***	19.321*
	(4.808)	(10.187)
T2 France	36.709***	36.509***
	(4.236)	(11.000)
T2 Japan	39.967***	40.893***
	(3.197)	(10.179)
T3 France	42.663***	42.363***
	(4.690)	(9.742)
T3 Japan	31.320***	31.411***
	(4.165)	(10.346)
Female		0.711
		(4.041)
Raven's Score		-0.053
		(0.678)
Economics		-0.049
& Management		(4.024)
Loss Aversion		-0.255
		(1.159)
Illusion of Control		0.268
		(0.246)
R^2	0.225	0.225
N	2740	2700

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regressions with clustered standard errors per subject.

differences (see Figure 5.D.3 in Appendix 5.D). Note that while principals in France attach on average a higher intrinsic value to be in control in T1 and T3 than in T2, principals in Japan attach the highest value in T2 and the lowest in T1. Table 5.3 shows that we reject the hypotheses that principals value the delegation lotteries and the control lotteries equally for any treatment in both countries.[31] This suggest that, on average, principals in France and Japan assign a positive intrinsic value to decision rights.

Several robustness tests bring support to Result 1. First, we observe that a large majority of principals derive a positive intrinsic value of holding control in all treatments and in both countries. In our total sample, 89% of principals value (on average) the delegation lotteries strictly more than the corresponding control lotteries. Second, a bootstrap analysis suggests that a positive and significant IV holds for all separate games (rounds) in France, and most games in Japan.[32] This suggests that a positive intrinsic value of decision rights is a robust preference across the different delegation games for most treatments. Finally, we test whether the intrinsic value of decision rights is measured *consistently* across principals in the ten delegation games.[33] To test for consistency, we follow Bartling et al. (2014) and measure the correlation of the IV across games by computing *Cronbach's alpha* (Cronbach 1951), in our case per treatment and country. The measure reports the correlation between the games, and varies between zero and one. Cronbach's alphas per treatment are between 0.49 (in T1) and 0.75 (in T3) in France and between 0.47 (in T2) and 0.60 (in T1) in Japan. This suggests a moderate and positive correlation of our IV measure across principals in France and Japan in the ten delegation games and all treatments.

[31] Nonparametric Wilcoxon signed-rank tests also reject this hypotheses ($p < 0.001$ for all cases), and similarly with bootstrap tests ($p < 0.001$ for all cases).

[32] For principals in France, bootstrap tests reject the hypothesis that the IV is equal to zero for 10 out of 10 games in T1 ($p < 0.01$ for 7, $p < 0.05$ for 2, and $p < 0.1$ for 1), for 10 games in T2 ($p < 0.01$ for all games), and for the 10 games in T3 ($p < 0.01$ for 9 and $p < 0.05$ for 1). For principals in Japan, bootstrap tests reject this hypothesis for 7 out of 10 games in T1 ($p < 0.01$ for 2 and $p < 0.05$ for 5), for 9 games in T2 ($p < 0.01$ for 8 and $p < 0.05$ for 1), and for 8 games in T3 ($p < 0.01$ for 7 and $p < 0.1$ for 1).

[33] Consistency means that if a principal assigns a higher intrinsic value to decision rights than another principal in one game, then the former also assigns a higher value in the other games.

Taken together, these results suggest that our first finding is robust for both countries and all treatments.

Result 2. *Decision rights have on average a higher intrinsic value for French than for Japanese principals.*

The positive IV that we observe in T1 for the two countries may be due, as exposed in Section 5.4, to any of the three rationales that we have identified. Then, the difference between countries of this measured value provides an estimation of the difference of IV taking all rationales into account. On average, the IV in T1 in France is worth 22 points more than the IV in Japan. We reject the hypothesis that principals in France and principals in Japan attach the same difference of value between the delegation lotteries and the control lotteries in T1 ($p < 0.001$, Wald test based on the regression of Table 5.3 with individual controls).[34]

5.5.3 The Roots of the Intrinsic Value of Decision Rights

The measured IV in our three treatments and their differences provide measures of our three main variables of interest: independence, power, and self-reliance. We present the measured values of these rationales in Table 5.4.[35] For instance, consider independence (R1) for principals in France. On average, principals attach 1.1 points less to the delegation lottery compared to the control lottery when they interact with a human agent (T1) than when they interact with a bot that decides on behalf of a passive human agent (T3).

Result 3. *Self-reliance is a significant rationale for the intrinsic value of decision rights for French and Japanese principals.*

As shown in the table, self-reliance seems to be a significant and positive rationale behind the IV in both countries. On average, the delegation lotteries are valued

[34] When taking the mean IV, a Mann-Whitney U test and a bootstrap test also reject this hypothesis ($p < 0.001$ and $p < 0.01$ respectively).

[35] See Table 5.B.2 in Appendix 5.B for the results of nonparametric tests and the robustness of the results.

Table 5.4 – The Roots of the Intrinsic Value of Decision Rights

	R1: Independence $(IV_{T1} - IV_{T3})$	R2: Power $(IV_{T3} - IV_{T2})$	R3: Self-reliance (IV_{T2})
France	-1.079	5.854	36.509^{***}
Japan	-12.090^{*}	-9.483^{*}	40.894^{***}
H_0: FR = JP	$p = 0.221$	$p = 0.066$	$p = 0.481$

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on two-sided Wald tests from OLS regression with clustered standard errors per subject and individual controls.

36.5 (in France) to 40.9 (in Japan) points more than the control lotteries due to a preference for self-reliance. As the Table 5.4 highlights, we reject the hypothesis that principals in T2 value the delegation lotteries and the control lotteries equally, both in France and Japan. Looking at the difference between the two countries, we do not reject the hypothesis that principals in France and principals in Japan attach the same difference of value between the delegation lotteries and the control lotteries in T2, i.e., we do not reject the hypothesis that $R3(France) = R3(Japan)$ ($p = 0.481$, two-tailed Wald test). This suggests that, on average, there is no significant difference in the value of self-reliance as a rationale behind the IV between France and Japan.

Result 4. *Independence is not a rationale for the intrinsic value of decision rights for French and Japanese principals.*

As it can be seen from Figure 5.3 and Table 5.4, French subjects tend to value holding control in T1 similarly than in T3, and we do not reject the null hypothesis of equality. This suggests that principals might be indifferent to independence as a rationale for the intrinsic value of decision rights in France. For Japan, independence has on average a negative impact on the IV. But as it can be seen from Table 5.4, we reject the hypothesis that independence has no effect on the IV only at 10% significance level. This brings then only mild evidence that independence might be negatively valued when it comes to be a motive of the intrinsic value of holding control in our setting. When comparing the measured values for the two countries in terms of independence, we do not reject the null hypothesis that $R1(France) = R1(Japan)$

($p = 0.221$, two-tailed Wald test). That is, though there is a slight difference in size, independence is similarly valued in France and in Japan as a rationale for the intrinsic value of holding control.

Result 5. *Power is not a rationale for the intrinsic value of decision rights for French and Japanese principals.*

As for power, principals give on average a positive value to this motive in France. However, we do not reject the hypotheses that this value is equal to zero for French principals. Then, as it is the case with independence, this suggests that French principals might be indifferent to power as a rationale for the intrinsic value of decision rights in our setting. As for principals in Japan, power is negatively valued but we reject the null hypothesis only at 10% significance level. This brings again only mild evidence that principals in Japan might see power as a negative rationale for the intrinsic value of holding control. Looking at differences between France and Japan in terms of power, we reject the hypothesis that $R2(France) = R2(Japan)$ ($p = 0.066$, two-tailed Wald test). Taken together with the size difference of independence, these results bring mild evidence of a cultural difference with respect to these motivations in our setting.

These findings bring interesting insights into the role of different motives on the control and delegation of authority. First, and contrary to **Hypothesis 1**, independence from interference is not a significant root of the intrinsic value of decision rights in both countries (Result 4). Second, our results suggest that contrary to **Hypothesis 2**, power is not a significant root of the intrinsic value of decision rights (Result 5). If anything, independence and power *per se* seem to have a negative impact in the intrinsic value of holding control in Japan. Third, and in accordance with **Hypothesis 3**, self-reliance seems to be an economically and statistically significant rationale behind the IV in both France and Japan (Result 3). In addition, its value seems to be similar in both countries. Finally, contrary to **Hypothesis 4**, we find no significant difference in terms of the value of independence between the two

countries. Still, Japanese principals give on average a lower value to both independence and power (the motivations related with another person) as rationales for a preference for holding control than French principals.

Our main results are robust to different stake sizes (see Appendix 5.C), different definitions of cultural background (see Appendix 5.E), and alternative explanations based on reciprocity, preference reversals, or corner solutions (see Appendix 5.F). We test for the effect of different degrees of conflict between the principal and the agent in Appendix 5.C. Next, we discuss some of the potential worries and limitations of Bartling et al.'s (2014) experimental design and our treatments.

5.6 Discussion

The first potential worry about the experimental design is related with the potential effects of principals' social and risk preferences. Note, however, that the indifference point between the control and the delegation lotteries is endogenously chosen based on the principals' unobserved social and risk preferences (see also Bartling et al. 2014, 2022). These preferences will then enter similarly into the determination of the certainty equivalents of the two lotteries in Part 2. It follows that the measured IV, since it is based on the *difference* between the certainty equivalents of the two lotteries, is computed *after* these preferences have been taken into account. This means that even though there is no potential effect of social preferences in T2 (contrary to T1 and T3), such preferences should not *a priori* bias our results.[36]

A second worry is related with the potential effects of the beliefs about the agents' or bots' behaviors. For example, the delegation mechanism in the experiment where agents choose an effort could cue principals for more familiar "real-world" setups in which principals expect their agents to shirk after delegation (see also Bartling

[36]This is based on the weak assumptions that principal's utility from a delegation lottery is increasing in the probability of success and that the principal weakly prefers if the agent chooses project $\mathcal{P}$ (see Bartling et al. 2014, 2018-9). See also Bartling et al. (2014, 2022) for a discussion of the potential (but unlikely) effects of extreme forms of inequality aversion.

et al. 2014, footnote 36). This could lead principals to increase the minimum effort requirements beyond their optimal indifference point to avoid that delegation occurs. Another possibility would be that principals could believe that an altruistic agent would choose a higher effort level if the agent would be less constrained from the minimum effort requirement (see Falk and Kosfeld 2006). This could lead principals to decrease the minimum effort requirements below their optimal indifference point to favor altruism from agents. Since these two phenomena go in opposite directions, it seems difficult to test if they are present. But in our data they seem to either cancel each other or to not be present. In France, the average minimum effort requirement per treatment was 62.155 (T1), 65.171 (T2), and 64.878 (T3). We find no statistically significant difference between treatments based on an OLS regression controlling for individual characteristics (see Tables 5.A.3 and 5.A.4 in Appendix 5.A). In Japan, the average minimum effort requirement per treatment was 53.663 (T1), 63.606 (T2), and 56.103 (T3), and we find that the value in T2 is statistically significantly greater than in T1 and T3 but not statistically significantly different between T1 and T3 (see again Tables 5.A.3 and 5.A.4 in Appendix 5.A).[37]

Note, however, that for any of these last explanations to be valid principals should have misunderstood the delegation mechanism. In fact, the principal (taking into account her risk and social preferences) should set the optimal choice of the minimum effort requirement $\underline{e}$ irrespective of the agent's or bot's effort choice. Recall that there is no feedback until the end of the experiment and that delegation takes place if and only if $e \geq \underline{e}$, which means that the principal has control over the minimum effort that the agent needs to choose for her to delegate the decision right. The instructions and control questions were designed such that the logic of setting an optimal minimum effort requirement was clearly understood, and our robustness results on bounded

[37] Another possibility would be for principals to use a kind of mixed strategy in response to randomness in T2 and T3. If present, this should translate into higher standard deviation on the minimum effort requirement in T2 and T3 when compared to T1. However, in France the individual standard deviation of $\underline{e}$ among the 10 games is not statistically significantly different between T1 (mean of 18.068) and T2 (mean of 15.240) or T3 (mean of 15.255), and in Japan it is higher in T1 (20.241) than in T2 (14.685) and T3 (17.788).

rationality bring some support for the non-significance of explanations linked to the misunderstanding of the instructions. In this sense, the measured IV, that depends upon the principal's choice of the minimum effort requirement $\underline{e}$, should be independent of beliefs about the agent's or bot's effort choice. Similarly, though principals in Part 1 of T1 are faced with risk and uncertainty, while principals in Part 1 of T2 and T3 are faced only with risk, the independence on beliefs, if it holds, indicates that this should not be an issue for our treatment comparisons.

In terms of the belief about the agents' chosen project, this could, in principle, have an effect on our treatment comparisons. In particular, if principals believed that project alternative $\mathcal{P}$ was chosen by agents with positive probability in T1, then we would underestimate the intrinsic value of decision rights in T1 with respect to T2 and T3 (see Bartling et al. 2014, 2022-3 on how this could underestimate the IV in T1). In our experiment, the agents' project choices indicate that this would be a reasonable anticipation specially in France (agents chose project $\mathcal{P}$ 32% of games in France and 9% of games in Japan).[38] If principals anticipated these probabilities correctly, we would underestimate the value of IV and independence and the differences between France and Japan in these two measures. But if this would be the case, behaviorally it should translate into a lower $\underline{e}$ in T1 than in T2 and T3. However, we find no significant differences for $\underline{e}$ between treatments in France and we find that $\underline{e}$ is significantly higher in T2 than in T1 and T3 in Japan (see Table 5.A.4 in Appendix 5.A). The principals own choices of project also bring mild evidence on the focus of own gains (though they are silent in terms of anticipations). Principals chose the project that gives them the higher payoff 84% of games in France and 88% of games in Japan.[39] But as with the other beliefs, we cannot exclude with certainty that this belief has not played a role in our experiment.

[38] Note that we exclude from these calculations the data from games 5 and 10, where the payoffs are the same for the two players, as well as the passive (non-incentivized) agents that participated in T3.

[39] As before (and whenever similar calculations are presented below) we exclude from these results the data from games 5 and 10.

Another worry is related with the comparison between Part 1 and Part 2 of the experiment. For example, a changing attitude towards risk between the two parts could in principle explain the positive difference between the certainty equivalent of the delegation lottery and the control lottery. Results by Abdellaoui, Baillon, Placido and Wakker (2011) suggest that risk aversion may depend on the source of ambiguity. Since Part 1 is a game and Part 2 is a lottery task, they involve different sources of ambiguity that could be behind such an effect.

To probe the idea of the potential effect of unstable risk attitudes across tasks, consider that subjects are risk neutral in the delegation task (Part 1) but risk averse in the lottery task (Part 2). If they are assumed to have a Constant Relative Risk Aversion (CRRA) utility function [i.e., $u(x) = x^{1-\rho}/(1-\rho)$], then simulations indicate that the difference in certainty equivalents will indeed be positive for almost all games and for any parameter $\rho > 0$. However, in order to explain the average values found in our experiment ρ needs to take a value of 2 or higher. But this estimated value seems to be considerably high when compared with the existing literature. For example, Holt and Laury (2002) found that the decisions made by a majority of subjects over paired lottery choices could be rationalized by a CRRA utility function with ρ between -0.15 and 0.68. Alternatively, consider that subjects evaluate lotteries based on $w(p)x + (1 - w(p))y$ with $x > y > 0$, where the function $w(p)$ is a probability weighting function. Adopting the weighting function proposed by Prelec (1998) [i.e., $w(p) = exp\Big(-\beta(-\ln(p))^{\alpha}\Big)$], L'Haridon and Vieider (2016) estimated a value of α of about 0.8 for Japanese subjects and 0.7 for French ones, while their estimation of β was about 0.95 for both countries. Following these estimates, we should find a mean difference in certainty equivalents of 6 points in Japan and 9.5 in France, which is well below our results. These results suggest that this effect, if present, could result in an upward bias of our estimates of the intrinsic value of decision rights but that it could not rationalize the differences in certainty equivalents that we found. In addition, the potential changing risk attitudes associated with the different tasks of

Part 1 and Part 2 should not, at least in principle, affect our treatment comparisons and estimations of independence and power.

Another potential limitation of this design is its overly abstract setting. In particular, there is no "real" delegation taking place in Part 1. Principals do not experience making "real-effort" to implement their chosen project when they hold control, neither observe (or are aware) of the agent making real-effort to implement his project when they delegate. The introduction of real-effort could enhance the real sense of delegation, though some evidence suggests that stated-effort is a good proxy for real-effort (e.g. Bruggen and Strobel 2007). On the other hand, the introduction of real-effort and a real sense of delegation would imply the introduction of feedback, which would introduce a dimension of learning that is controlled in the current design.

Closely related to this issue is the absence of contextual cues to enhance a feeling of delegation in the experiment. Though the absence of contextual cues is worrisome for the understanding of how these preferences might adapt to "real-world" settings, it seems a reasonable methodological choice for cross-cultural comparisons. In fact, there is substantive evidence that contextual cues are interpreted differently by people from different cultural backgrounds (see e.g. Nisbett 2003; Nisbett and Masuda 2003). Then, it could become difficult to separate the effect of contextual cues from the effect of the IV in a less abstract setting. Still, it is also reasonable to expect that contextual cues that favor either feelings of independence or power could change the weight of these rationales in the IV. An important question is if these rationales would overcrowd the weight of self-reliance that we found in our abstract setting where the willingness to implement a decision (and/or choice) that a subject arrived at after some cognitive effort might be the crucial aspect of the intrinsic value of decision rights.

5.7 Conclusion

Several recent experimental analyses provide us with new insight on the incentive effects of decision rights and the preference for holding control. For instance, while principals often use control to reduce the agent's self-seeking actions, experimental results of Falk and Kosfeld (2006) suggest that holding control may carry some "hidden costs" in terms of agents' performance and principals' payoffs. In a context of participative decision making, Corgnet and Hernán-González (2014) show that consulting agents is beneficial for principals only if they follow the agent's choice. However, we have few insight into the motivations that lie behind the value of holding control when there is no instrumental reason to pursue it. This chapter is an attempt to shed some light about the motivations that lie behind the non-delegation of decision rights and its potential contextual determinants.

We find that, somewhat surprisingly, independence and power are not motivations that lie behind an intrinsic value to hold control in our setting. Instead, the will to apply one's decision seems to be the main motivation behind this preference. We find that these preferences are shared by French and Japanese subjects, but that the intrinsic value of decision rights is higher in France in good part because independence and power are on average negatively valued in Japan. So in general, while our results support the main finding of Bartling et al. (2014) concerning the positive and statistically significant intrinsic value of decision rights in this setting, they suggest that this value is not dependent upon another person being the potential holder of control.

Future research should investigate more thoroughly the effects of beliefs on this type of experimental settings. A potential way to do so for this experiment would be to run some additional sessions of T2 and/or T3 with a bot that would make choices consistent with agents that have previously participated in the experiment (e.g. either taking an average of agents' behavior or the behavior of one randomly chosen agent in particular). If beliefs play no fundamental role in our design, we

should find non-significant differences between these additional sessions and the sessions where the bot chooses project $\mathcal{A}$ and randomly determines its effort (as in the current design).

Finally, remark that the willingness to keep control in treatments 2 and 3 could, at least in principle, be due to motives linked to the chance device *per se* (e.g. aversion to randomness). Our findings would be consistent with aversion to randomness if one interprets R3 to be an aggregated preference for self-reliance *and* the willingness not to be affected by a chance device (nature). Our preferred interpretation is that subjects exhibit a preference for self-reliance despite the fact that they could delegate to a random device, since previous experiments suggest that subjects tend to "delegate" the decision to "randomness" in order to avoid some subjective costs (e.g. cognitive cost, responsibility aversion or regret; see Dwenger, Kübler and Weizsacker 2014 and Agranov and Ortoleva 2017). Future research could try to separate these two motives and shed further light on the effects of the interference of random devices on people's decisions to hold or delegate control.

All-in-all, the experimental setting is fairly abstract and also demanding in terms of understanding and rationality requirements. A companion study with an easier design and more contextual cues could help to understand how the intrinsic value of decision rights might translate to "real-world" settings. In such an experiment, one possible interesting extension would be to test for within-country and between-country variation taking different subject-pools (say students and factory workers.). This would allow, among other things, to test if the variation in the roots of the intrinsic value of decision rights is greater within a country than the variation between two countries.

Bibliography

Abdellaoui, M., A. Baillon, L. Placido, and P. P. Wakker (2011) The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation. The American Economic Review 101(2): 695–723.

Aghion, P. and P. Bolton (1992) An Incomplete Contracts Approach to Financial Contracting. Review of Economic Studies 59(3): 473–94.

Aghion, P. and J. Tirole (1997) Formal and Real Authority in Organizations. Journal of Political Economy 105(1): 1–29.

Agranov, M. and P. Ortoleva (2017) Stochastic Choice and Preferences for Randomization. Journal of Political Economy 125(1): 40–68.

Anderson, J. (2013) Autonomy. In: H. LaFollette (ed) The International Encyclopedia of Ethics. Blackwell Publishing Ltd., first edition edition.

Bartling, B., E. Fehr, and H. Herz (2014) The Intrinsic Value of Decision Rights. Econometrica 82(6): 2005–39.

Bartling, B. and U. Fischbacher (2012) Shifting the Blame: On Delegation and Responsibility. Review of Economic Studies 79(1): 67–87.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964) Measuring Utility by a Single-Response Sequential Method. Behavioral Science 9: 226–32.

Berg, J. E., J. W. Dickhaut, and T. A. Rietz (2010) Preference reversals: The impact of truth-revealing monetary incentives. Games and Economic Behavior 68: 443–468.

Berlin, I. (1969) Four Essays on Liberty. Oxford University Press., Oxford.

Bohnet, I., F. Greig, B. Herrmann, and R. Zeckhauser (2008) Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. The American Economic Review 98(1): 294–310.

Bowles, S. and H. Gintis (1988) Contested Exchange: Political Economy and Modern Economic Theory. The American Economic Review: Papers and Proceedings 78(2): 145–50.

Bruggen, A. and M. Strobel (2007) Real Effort Versus Chosen Effort in Experiments. Economic Letters 96: 232–6.

Buchanan, J. M. (1986) Liberty, Market and the State. Wheatsheaf Books, Brighton.

Burdin, G., S. Halliday, and F. Landini (2015) Third-Party vs. Second-Party Control: Disentangling the Role of Autonomy and Reciprocity. , IZA Working Paper 9251.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009) Cognitive skills affect economic preferences, strategic behavior, and job attachment. Proceedings of the National Academy of Science, U.S.A. 106(19): 7745–7750.

Carpenter, J., M. Graham, and J. Wolf (2013) Cognitive ability and strategic sophistication. Games and Economic Behavior 80: 115–130.

Charness, G. and U. Gneezy (2010) Portfolio Choice and Risk Attitudes: An Experiment. Economic Inquiry 48(1): 133–46.

Christman, J. (2011) Autonomy in Moral and Political Philosophy. http://plato.stanford.edu/archives/spr2011/ entries/autonomy-moral/.

Corgnet, B. and R. Hernán-González (2014) Don't ask me if you will not listen: The dilemma of participative decision making. Management Science 60(3): 560–585.

Cronbach, L. J. (1951) Coefficient Alpha and the Internal Structure of the Tests. Psychometrika 16: 297–334.

Dahl, R. A. (1957) The Concept of Power. Behavioral science 2(3): 201–15.

Deci, E. L. and R. M. Ryan (1985) Intrinsic Motivation and Self-Determination in Human Behavior. Plenum, New York.

Dwenger, N., D. Kübler, and G. Weizsacker (2014) Flipping a Coin: Theory and Evidence. , Working Paper.

Dyck, A. and L. Zingales (2004) Private Benefits of Control: An International Comparison. Journal of Finance 59(2): 537–600.

Falk, A. and M. Kosfeld (2006) The Hidden Costs of Control. The American Economic Review 96: 1611–30.

Fehr, E. and L. Goette (2007) Do Workers Work More if Wages Are High? Evidence From a Randomized Field Experiment. The American Economic Review 97(1): 298–317.

Fehr, E., H. Herz, and T. Wilkening (2013) The Lure of Authority: Motivation and Incentive Effects of Power. The American Economic Review 103(4): 1325–59.

Fehr, E. and K. Hoff (2011) Introduction: Tastes, Castes and Culture: The Influence of Society on Preferences. The Economic Journal 211: 396–412.

Fischbacher, U. (2007) z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. Experimental Economics 10(2): 171–8.

Gachter, S., B. Herrmann, and C. Thoni (2010) Culture and Cooperation. Philosophical Transactions of the Royal Society B 265: 2651–61.

Gill, D. and V. Prowse (2016) Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-k Analysis. Journal of Political Economy 124(6): 1619–1676.

Grether, D. M. and C. R. Plott (1979) Economic Theory of Choice and the Preference Reversal Phenomenon. American Economic Review 69(4): 623–638.

Hamilton, B. H. (2000) Does Entrepreneurship Pay? An Empirical Analysis of the Returns to Self-Employment. Journal of Political Economy 108(3): 604–31.

Hanaki, N., N. Jacquemet, S. Luchini, and A. Zylbersztejn (2016a) Cognitive ability and the effect of strategic uncertainty. Theory and Decision 81(1): 101–121.

——— (2016b) Fluid intelligence and cognitive reflection in a strategic environment: evidence from dominance-solvable games. Frontiers in Psychology: Personality and Social Psychology 10(http://dx.doi.org/10.3389/fpsyg.2016.01188).

Hart, O. and J. Moore (1990) Property Rights and the Nature of the Firm. Journal of Political Economy 98(6): 1119–58.

——— (1995) Debt and Seniority: An Analysis of the Role of Hard Claims in Constraining Management. The American Economic Review 85(3): 567–85.

Hayek, F. A. (1960) The Constitution of Liberty. Routledge., London.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis (2004) Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence From Fifteen Small-scale Societies. Oxford University Press, Oxford.

Henrich, J., J. Ensminger, R. McElreath, A. Barr, C. Barrett, A. Bolyanatz, J. Camilo Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, and J. Ziker (2010) Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. Science 327: 1480–4.

Holt, C. R. and S. K. Laury (2002) Risk aversion and Incentive Effects. The American Economic Review 92(5): 1644–55.

Inglehart, R. and W. E. Baker (2000) Modernization, Cultural Change, and the Persistence of Traditional Values. American Sociological Review 65(1): 19–51.

Iyengar, S. S. and S. E. DeVoe (2003) Rethinking the Value of Choice: Considering Cultural Mediators of Intrinsic Motivation. In: V. Murphy-Berman and J. J. Berman (eds) Nebraska symposium on motivation: Cross-cultural differences in perspectives on selfV. Murphy-Berman and Self-Regulation and Human Autonomy 1583. 49 University of Nebraska Press, Lincoln: 129–74.

Iyengar, S. S. and M. R. Lepper (1999) Rethinking the Value of Choice: A Cultural Perspective on Intrinsic Motivation. Journal of Personality and Social Psychology 76(3): 349–66.

LeBoeuf, R. A., E. Shafir, and J. B. Bayuk (2010) The Conflicting Choices of Alternating Selves. Organizational Behavior and Human Decision Processes 111(1): 48–61.

Lee, F. (1997) When the Going Gets Tough, Do the Tough Ask for Help? Help Seeking and Power Motivation in Organizations. Organizational Behavior and Human Decision Processes 72(3): 336–63.

L'Haridon, O. and F. M. Vieider (2016) All over the map: A Worldwide Comparison of Prospect Theory Parameters. , Working Paper.

Lichtenstein, S. and P. Slovic (1971) Reversals of preference between bids and choices in gambling decisions. Journal of Experimental Psychology 89(1): 46–55.

Lynn, R. and P. Irwing (2004) Sex Differences on the Progressive Matrices: A Meta-analysis. Intelligence 32: 481–98.

Machiavelli, N. (2003) The Discourses. Penguin Classics.

Markus, H. R. and S. Kitayama (1991) Culture and the Self: Implications for Cognition, Emotion, and Motivation. Psychological Review 98: 224–53.

McClelland, D. C. (1975) Power: The Inner Experience. Irvington. Halsted Press, Irvington, NY.

Neri, C. and H. Rommeswinkel (2014) Freedom, Power and Interference. , University of St. Gallen.

Nisbett, R. E. (2003) The Geography of Thought: How Asians and Westerners Think Differently...and Why. The Free Press, New York.

Nisbett, R. E. and T. Masuda (2003) Culture and Point of View. Proceedings of the National Academy of Sciences of the United States of America 100: 11163–75.

Nozick, R. (1974) Anarchy, State, and Utopia. Basic Books.

Owens, D., Z. Grossman, and R. Fackler (2014) The Control Premium: A Preference for Payoff Autonomuy. American Economic Journal: Microeconomics 6(4): 138–61.

Parks, C. D. and A. D. Vu (1994) Social Dilemma Behavior of Individuals From Highly Individualist and Collectivist Cultures. Journal of Conflict Resolution 38: 708–18.

Patton, P. (1989) Taylor and Foucault on Power and Freedom. Political Studies 352: 260–76.

Prelec, D. (1998) The Probability Weighting Function. Econometrica 66: 497–527.

Pugsley, B. W. and E. Hurst (2011) What Do Small Businesses Do?. Brookings Papers on Economic Activity 43(2): 73–142.

Raven, J. C. (1998) Raven's Advanced Progressive Matrices (APM). Pearson, San Antonio, TX2003rd edition.

Rawls, J. (1971) A Theory of Justice. Harvard University Press, Cambridge.

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

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991) Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. The American Economic Review 81(5): 1068–95.

Ryan, R. M. and E. L. Deci (2006) Self-Regulation and the Problem of Human Autonomy: Does Psychology Need Choice, Self-Determination, and Will?. Journal of Personality 74(6): 1557–86.

Sen, A. K. (1988) Freedom of Choice: Concept and Content. European Economic Review 32: 269–94.

Simon, H. A. (1951) A Formal Theory of the Employment Relation. Econometrica 19(3): 293–05.

Skinner, Q. (1992) On Justice, the Common Good and the Priority of Liberty. In: C. Mouffe (ed) Dimensions of Radical Democracy: Pluralism, Citizenship, Community. Verso, London.

Welzel, C. and R. Inglehart (2005) Liberalism, Postmaterialism, and the Growth of Freedom. International Review of Sociology 15(1): 81–108.

Wong, R. Y.-M. and Y.-Y. Hong (2005) Dynamic Influences of Culture on Cooperation in the Prisoner's Dilemma. Psychological Science 16: 429–34.

Appendix

5.A Decisions Part 1

Table 5.A.1 – Within Country Differences, $\underline{e}$

	Treatment 1		Treatment 2		Treatment 3	
Rennes	60.575***	55.650***	63.842***	66.229***	61.629***	70.919***
	(2.829)	(9.809)	(2.078)	(6.166)	(2.612)	(9.572)
Nice	62.733***	58.604***	65.660***	67.775***	70.046***	77.194***
	(5.204)	(11.036)	(2.186)	(6.347)	(5.394)	(9.456)
Tsukuba	54.025***	46.453***	61.108***	67.532***	56.882***	71.325***
	(3.968)	(10.948)	(2.766)	(6.871)	(2.774)	(11.166)
Osaka	49.791***	43.735***	60.667***	65.893***	52.862***	62.664***
	(3.709)	(10.538)	(3.034)	(7.651)	(4.574)	(10.230)
Female		−4.095		−1.653		0.880
		(4.114)		(2.875)		(3.902)
Raven's Score		0.650		−0.322		−0.821
		(0.727)		(0.466)		(0.748)
Economics & Management		−0.152		4.934*		4.705
		(4.335)		(2.759)		(3.820)
Loss Aversion		0.111		−0.597		−1.727
		(0.989)		(0.803)		(1.309)
Illusion of Control		0.059		0.175		0.143
		(0.280)		(0.186)		(0.274)
R^2	0.797	0.800	0.889	0.894	0.836	0.840
N	860	850	1010	990	870	860
H_0: Rennes = Nice	0.717	0.598	0.548	0.608	0.164	0.315
H_0: Tsukuba = Osaka	0.438	0.644	0.915	0.689	0.455	0.147

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regressions with clustered standard errors per subject.

Table 5.A.2 – Within Country Differences, E

	Treatment 1		Treatment 2		Treatment 3	
Rennes	70.682***	68.304***	63.926***	60.512***	68.950***	75.596***
	(2.350)	(9.500)	(2.538)	(6.570)	(2.662)	(9.485)
Nice	54.213***	52.821***	64.455***	61.629***	67.646***	73.220***
	(4.988)	(10.589)	(2.809)	(6.820)	(5.008)	(9.450)
Tsukuba	63.710***	60.485***	63.000***	62.812***	65.318***	77.271***
	(3.621)	(9.985)	(2.541)	(7.793)	(2.054)	(11.276)
Osaka	55.400***	52.22***	62.111***	60.757***	59.367***	67.306***
	(2.825)	(9.344)	(3.171)	(7.883)	(3.856)	(9.969)
Female		−0.570		−1.200		6.120*
		(3.913)		(3.329)		(3.282)
Raven's Score		0.314		−0.056		−0.814
		(0.618)		(0.538)		(0.745)
Economics & Management		−2.946		3.999		0.271
		(3.692)		(3.395)		(3.430)
Loss Aversion		0.123		0.688		−1.018
		(0.851)		(1.092)		(1.107)
Illusion of Control		−0.062		−0.016		−0.008
		(0.234)		(0.248)		(0.227)
R^2	0.875	0.875	0.896	0.900	0.888	0.892
N	860	850	1010	990	870	860
H_0: Rennes = Nice	0.004	0.003	0.889	0.775	0.819	0.668
H_0: Tsukuba = Osaka	0.074	0.106	0.827	0.620	0.177	0.043

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regressions with clustered standard errors per subject.

Table 5.A.3 – Minimum Effort Requirement, $\underline{e}$

T1 France	61.328***	62.155***
	(2.579)	(5.855)
T1 Japan	51.760***	53.663***
	(2.716)	(5.816)
T2 France	64.469***	65.171***
	(1.555)	(4.874)
T2 Japan	60.923***	63.606***
	(2.042)	(5.522)
T3 France	64.298***	64.878***
	(2.535)	(5.040)
T3 Japan	54.785***	56.103***
	(2.734)	(5.912)
Female		-1.979
		(2.062)
Raven's Score		-0.003
		(0.379)
Economics & Management		3.698*
		(2.028)
Loss Aversion		-0.649
		(0.595)
Illusion of Control		0.096
		(0.157)
R^2	0.844	0.846
N	2740	2700

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regression with clustered standard errors per subject.

Table 5.A.4 – Minimum Effort Requirement, $\underline{e}$, Between Treatement Differences

France	Wald Tests
H_0: T1 = T2	0.328
H_0: T1 = T3	0.453
H_0: T2 = T3	0.921
Japan	
H_0: T1 = T2	0.004
H_0: T1 = T3	0.529
H_0: T2 = T3	0.030

Notes: This table displays p-values for two-tailed Wald tests applied on the minimum effort requirement $\underline{e}$ per subject and game with individual controls and clustered standard errors per subject.

Table 5.A.5 – Effort, E

T1 France	64.937***	61.006***
	(2.599)	(6.060)
T1 Japan	59.265***	56.804***
	(2.339)	(5.479)
T2 France	64.109***	60.984***
	(1.917)	(5.025)
T2 Japan	62.628***	60.986***
	(1.980)	(5.484)
T3 France	68.537***	65.608***
	(2.404)	(5.071)
T3 Japan	62.213***	59.941***
	(2.271)	(5.777)
Female		1.4957
		(2.045)
Raven's Score		-0.085
		(0.369)
Economics & Management		0.811
		(2.047)
Loss Aversion		0.263
		(0.579)
Illusion of Control		-0.003
		(0.150)
R^2	0.884	0.885
N	2740	2700

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regression with clustered standard errors per subject.

Table 5.A.6 – Effort, E, Between Treatement Differences

France	Wald Tests
H_0: T1 = T2	0.995
H_0: T1 = T3	0.203
H_0: T2 = T3	0.132
Japan	
H_0: T1 = T2	0.177
H_0: T1 = T3	0.333
H_0: T2 = T3	0.731

Notes: This table displays p-values for two-tailed Wald tests applied on the effort E per subject and game with individual controls and clustered standard errors per subject.

5.B Nonparametric Tests

Table 5.B.1 – Within Country Differences:
Parametric and non parametric tests

	Treatment 1	Treatment 2	Treatment 3
H_0: Rennes = Nice			
Wilcoxon-Mann-Whitney test	0.302	0.432	0.867
Student test	0.209	0.154	0.661
H_0: Tsukuba = Osaka			
Wilcoxon-Mann-Whitney test	0.609	0.844	0.482
Student test	0.744	0.865	0.703

Notes: This table displays p-values for two-tailed Wilcoxon-Mann-Whitney and Student tests applied on the average IV by subject.

Table 5.B.2 – The Roots of the Intrinsic Value of Decision Rights (Nonparametric Tests)

	R1: Independence $(IV_{T1} - IV_{T3})$	R2: Power $(IV_{T3} - IV_{T2})$	R3: Self-reliance (IV_{T2})
France	-0.394	5.955	36.709^{***}
Japan	-12.387^{*}	-8.648^{*}	39.967^{***}

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on Mann-Whitney U tests (R1 and R2) and Wilcoxon signed-rank test (R3) and with average IV by subject.

5.C Situational Determinants

The experimental design allows us to test for two situational determinants of the IV and its rationales: (i) stake size and (ii) conflict of interest. In terms of stake size, it is possible to distinguish between "low stakes" games (1-5) and "high stakes" games (6-10) [see Table 5.2].

Table 5.C.1 shows the intrinsic value of decision rights and its rationales for the two stake size levels. As it can be seen from the Table, the magnitudes of all our estimated absolute values increase with the stake size. In terms of the IV, this is consistent with the findings of Bartling et al. (2014). One can note that this is not surprising, since the percentage difference also increases with the stake size. The IV is positive and significantly different between France and Japan for both low and high stake sizes. This is again consistent with our main results. The only difference with respect to our main analysis is that the IV in Japan becomes insignificant for low stakes. But either using percentage difference as the dependent variable or non-parametric tests with average IV by individual the intrinsic value of holding control is highly significant for low stakes also in Japan ($p = 0.019$ based on OLS regression with clustered standard errors per subject and individual controls, and $p = 0.003$ based on a Wilcoxon signed-rank test respectively). In terms of the roots of the intrinsic value of holding control, our main results are robust to the stake size. In particular, self-reliance continues to be the only positive and significant rationale of the IV. In addition, independence and power continue to be non-significantly valued in France and on average negatively valued (though either not significantly or significantly just at 10%) in Japan.

In terms of the conflict of interest between the principal and the agent, Bartling et al. (2014) distinguish between "high conflict" games (3, 4, 8 and 9), "low conflict" games (1, 2, 6 and 7), and "no conflict" games (5 and 10) [see Table 5.2].[40] However, we cannot do the same type of analysis as in Table 5.C.1 given that the stake size is

[40] In Bartling et al. (2014) conflict of interest is defined as the principal's *relative* payoff difference between projects $\mathcal{A}$ and $\mathcal{P}$ $[(P_{\mathcal{A}} - P_0) \ / \ (P_{\mathcal{P}} - P_0)]$.

Table 5.C.1 – The Effect of Stake Size

	IV IV_{T1}	R1: Independence $(IV_{T1} - IV_{T3})$	R2: Power $(IV_{T3} - IV_{T2})$	R3: Self-reliance (IV_{T2})
France (Low)	29.767^{***}	1.351	2.450	25.966^{***}
France (High)	52.800^{***}	-3.509	9.257	47.052^{***}
H_0: Low = High	$p = 0.001$	$p = 0.565$	$p = 0.395$	$p = 0.001$
Japan (Low)	11.965	-8.166^{*}	-4.355	24.486^{**}
Japan (High)	26.677^{**}	-16.014	-14.610^{*}	57.301^{***}
H_0: Low= High	$p = 0.026$	$p = 0.395$	$p = 0.240$	$p < 0.001$
H_0: FR = JP (Low)	$p < 0.001$	$p = 0.141$	$p = 0.272$	$p = 0.769$
H_0: FR = JP (High)	$p = 0.014$	$p = 0.375$	$p = 0.069$	$p = 0.281$

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regression with clustered standard errors per subject and individual controls.

different for different degrees of conflict. In accordance, it would not be possible to disentangle the two effects. In order to further investigate our results concerning different stake sizes and have a sense of the effect of the degree of conflict between the principal and the agent, we test for the marginal effects of the stake size and degree of conflict of interest on our IV measure.

Table 5.C.2 shows the marginal effect of the stake size and degree of conflict on the IV, per treatment and country.[41] As seen from the Table, and taking all rationales into account (i.e., values for T1), while the intrinsic value of holding control increases 0.155 and 0.075 points per additional unit of stake size in France and Japan respectively, it decreases 0.341 and 0.128 points for each additional unit of degree of conflict in France and Japan respectively. This is consistent with the findings of Bartling et al. (2014) in Switzerland. In terms of the roots of the intrinsic value of holding control, the marginal effect of stake size is only significant for self-reliance in France ($p < 0.001$, Wald test based on the OLS regression of Table 5.C.2 with in-

[41] The principal's payoff of project A in case of success (P_A) is used as a proxy for the stake size of each game while the difference between the principal's payoff and the agent's payoff of project A in case of success ($P_A - A_A$) is used as a proxy for the degree of conflict in each game. We use project A since it determines the high payoff of the delegation lottery for which the principal has not to pay the cost of effort.

Table 5.C.2 – The Marginal Effect of Stake Size and Conflict of Interest

T1 France		14.153	13.393
		(9.783)	(13.564)
T1 Japan		3.345	3.733
		(7.636)	(11.867)
T2 France		9.848	9.649
		(6.394)	(9.753)
T2 Japan		-5.503	-2.786
		(6.851)	(11.163)
T3 France		10.462*	10.162
		(6.002)	(10.504)
T3 Japan		-3.026	-1.044
		(8.035)	(11.668)
Stake*T1 France		0.156***	0.155***
		(0.034)	(0.035)
Stake*T2 France		0.125***	0.125***
		(0.030)	(0.030)
Stake*T3 France		0.168***	0.168***
		(0.026)	(0.026)
Stake*T1 Japan		0.075**	0.075**
		(0.036)	(0.036)
Stake*T2 Japan		0.209***	0.207***
		(0.030)	(0.031)
Stake*T3 Japan		0.171***	0.171***
		(0.034)	(0.035)
Conflict*T1 France		-0.344***	-0.341***
		(0.120)	(0.123)
Conflict*T2 France		-0.192**	-0.192**
		(0.077)	(0.077)
Conflict*T3 France		-0.331***	-0.331***
		(0.092)	(0.092)
Conflict*T1 Japan		-0.128	-0.128
		(0.121)	(0.121)
Conflict*T2 Japan		-0.309***	-0.334***
		(0.077)	(0.076)
Conflict*T3 Japan		-0.308***	-0.342***
		(0.088)	(0.083)
Female			0.711
			(4.050)
Raven's Score			-0.053
			(0.679)
Economics & Management			-0.049
			(4.033)
Loss Aversion			-0.256
			(1.162)
Illusion of Control			0.268
			(0.246)
R^2	278	0.299	0.300
N		2740	2700

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regressions with clustered standard errors per subject.

dividual controls), and for self-reliance and independence in Japan ($p < 0.001$ and $p = 0.060$ respectively, Wald test based on the OLS regression of Table 5.C.2 with individual controls). As for the marginal effect of degree of conflict, it is only significant for self-reliance in France and Japan ($p = 0.013$ and $p < 0.001$ respectively, Wald test based on the OLS regression of Table 5.C.2 with individual controls). Though this is surprising, it seems to accord with the result that in our setting independence and power are not behind the intrinsic value of decision rights.

5.D IV in Percentage Difference

Table 5.D.1 – Within Country Differences, IV/CE

	Treatment 1		Treatment 2		Treatment 3	
Rennes	0.246***	0.267***	0.187***	0.185*	0.187***	0.239***
	(0.035)	(0.082)	(0.034)	(0.097)	(0.025)	(0.073)
Nice	0.190***	0.215**	0.239***	0.248**	0.231***	0.280***
	(0.034)	(0.087)	(0.034)	(0.099)	(0.040)	(0.069)
Tsukuba	0.100**	0.131	0.177***	0.205**	0.138***	0.217***
	(0.041)	(0.093)	(0.018)	(0.095)	(0.022)	(0.078)
Osaka	0.090***	0.121	0.191***	0.201**	0.138***	0.192***
	(0.028)	(0.073)	(0.029)	(0.094)	(0.026)	(0.070)
Female		0.020		0.022		0.021
		(0.040)		(0.039)		(0.029)
Raven's Score		−0.005		−0.005		−0.004
		(0.005)		(0.007)		(0.006)
Economics & Management		−0.031		−0.008		−0.010
		(0.039)		(0.043)		(0.031)
Loss Aversion		0.006		0.012		−0.010
		(0.011)		(0.015)		(0.008)
Illusion of Control		0.001		0.000		0.001
		(0.002)		(0.003)		(0.002)
R^2	0.225	0.223	0.270	0.276	0.292	0.306
N	860	850	1010	990	870	860
H_0: Rennes = Nice	0.252	0.303	0.282	0.172	0.355	0.412
H_0: Tsukuba = Osaka	0.834	0.867	0.664	0.918	0.991	0.480

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regressions with clustered standard errors per subject.

Table 5.D.2 – The Intrinsic Value of Decision Rights, IV/CE

T1 France	0.226***	0.250***
	(0.026)	(0.053)
T1 Japan	0.095***	0.133**
	(0.024)	(0.051)
T2 France	0.205***	0.233***
	(0.026)	(0.057)
T2 Japan	0.183***	0.223***
	(0.016)	(0.050)
T3 France	0.201***	0.230***
	(0.022)	(0.048)
T3 Japan	0.138***	0.174***
	(0.017)	(0.048)
Female		0.021
		(0.021)
Raven's Score		-0.004
		(0.004)
Economics & Management		-0.014
		(0.023)
Loss Aversion		0.002
		(0.007)
Illusion of Control		0.000
		(0.001)
R^2	0.258	0.261
N	2740	2700

Notes: *Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regressions with clustered standard errors per subject.

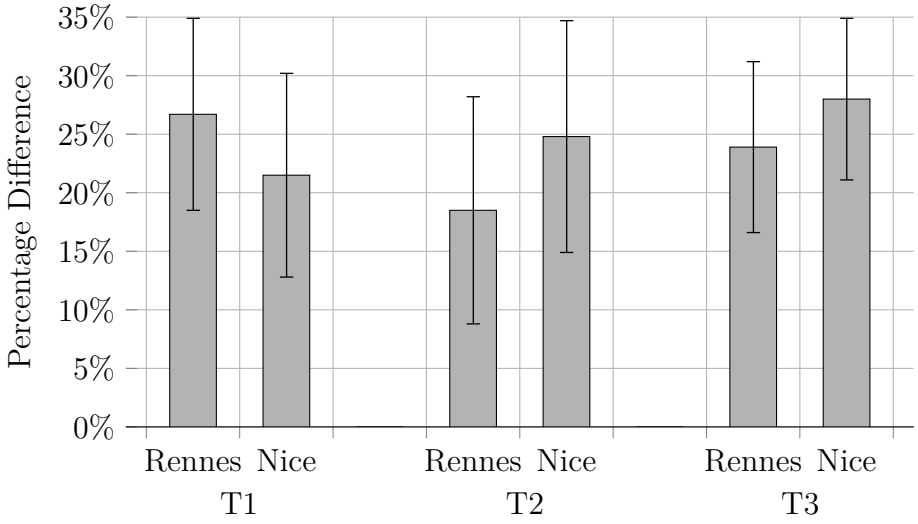


Figure 5.D.1 – IV/CE, sorted by French location and treatment. The bars display one standard error of the mean.

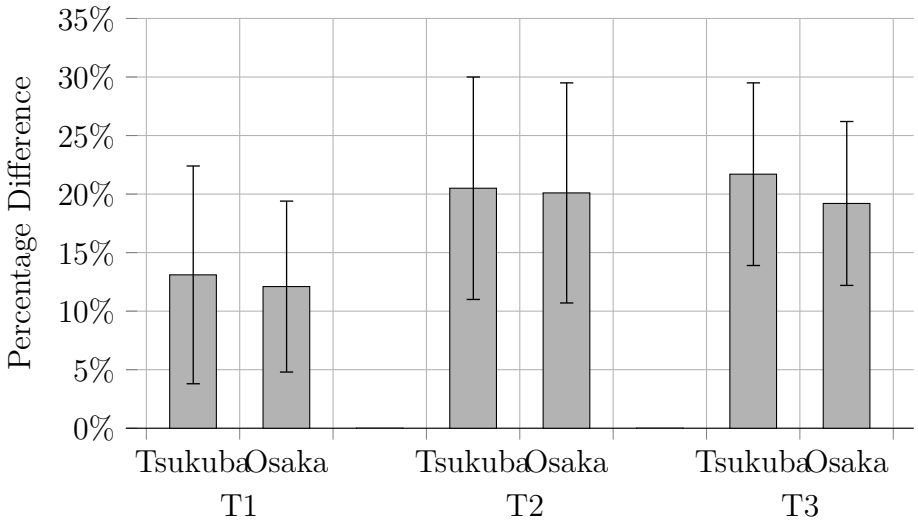


Figure 5.D.2 – IV/CE, sorted by Japanese location and treatment. The bars display one standard error of the mean.

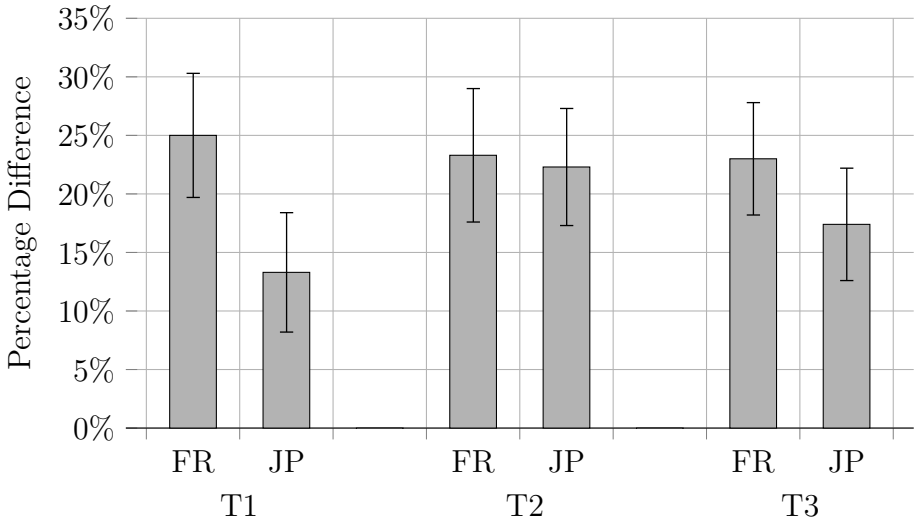


Figure 5.D.3 – IV/CE, sorted by country and treatment. The bars display one standard error of the mean.

5.E Robustness to Other Definitions of Cultural Background

Since we are interested in the effect of the cultural background, in the main analysis we have excluded subjects that are neither born in France or in Japan *and* are not of French or Japanese nationality. Our results are robust to other definitions of link to the country or cultural background, that either "weaken" (Definitions 1, 2, and 3) or "strengthen" (Definitions 4 and 5) the definition used in the main text. We test for the following definitions:

- **Definition 1**: A subject is said to be French/Japanese if she/he was born in France/Japan *or* one of her/his parents was born in France/Japan.
- **Definition 2**: A subject is said to be French/Japanese if she/he is of French/Japanese nationality.
- **Definition 3**: Union of Definition 1 and Definition 2.
- **Definition 4**: Intersection of Definition 1 and Definition 2.
- **Definition 5**: Intersection of Definition 1 and Definition 2 *and* at least one parent was born in France/Japan.

Table 5.E.1 reports the number of subjects for the different definitions. Table 5.E.2 presents the summary of the main results in terms of sign and significance, where "Def 0" represents the definition used in the main text.

Table 5.E.1 – Number of Subjects per Definition

	France			**Japan**	
Def 1 ↓ / Def 2 →	French	Not French	Def 1 ↓ / Def 2 →	Japanese	Not Japanese
French	133	9	Japanese	130	4
Not French	8	32	Not Japanese	0	2

For definition 5 remain 121 subjects in France and 130 subjects in Japan.

Table 5.E.2 – Summary of the Results per Definition

	France	Japan	FR - JP
R1: Independence			
Def 0:	0	$-^{*}$	0
Def 1:	0	$-^{*}$	$+^{*}$
Def 2:	0	$-^{**}$	0
Def 3:	0	$-^{*}$	0
Def 4:	0	$-^{**}$	$+^{*}$
Def 5:	0	$-^{**}$	$+^{**}$
R2: Power			
Def 0:	0	$-^{*}$	$+^{*}$
Def 1:	0	$-^{*}$	0
Def 2:	0	$-^{*}$	$+^{*}$
Def 3:	0	$-^{*}$	0
Def 4:	0	$-^{*}$	$+^{*}$
Def 5:	0	$-^{*}$	$+^{*}$
R3: Self-reliance			
Def 0:	$+^{***}$	$+^{***}$	0
Def 1:	$+^{***}$	$+^{***}$	0
Def 2:	$+^{***}$	$+^{***}$	0
Def 3:	$+^{***}$	$+^{***}$	0
Def 4:	$+^{***}$	$+^{***}$	0
Def 5:	$+^{***}$	$+^{***}$	0

*Significant at 10% level, **Significant at 5% level, ***Significant at 1% level, based on OLS regressions with clustered standard errors per subject and individual controls. The null hypothesis for the differences between France and Japan are on two-sided tests (H_0: FR = JP).

5.F Alternative Explanations

These findings are robust to alternative explanations based on loss aversion, illusion of control, or bounded rationality discussed in Bartling et al. (2014). Our regressions show that these have no significant impact on the intrinsic value of decision rights. Our data is neither consistent with explanations based on reciprocity, preference reversals, or corner solutions (see Bartling et al. 2014 for details). If reciprocity would be behind the measured IV, the differences in the certainty equivalents between the delegation and control lotteries in Part 2 should be higher the lower the minimum effort requirement imposed by the principal in Part 1. However, the data do not lend support to this trend. In a regression of the percentage difference in certainty equivalents on the minimum effort requirement, controlling for subject and game fixed effects, the percentage difference in the certainty equivalents increases by 3.5 percentage points per 10 point increase in the minimum effort requirement ($p < 0.001$, standard errors clustered at the subject level).

In terms of preference reversals, there exists a large literature showing that people tend to overbid a high-amount lottery in a pricing task (as the Part 2 of our experiment is) while preferring a high-probability lottery in a binary choice (e.g. Lichtenstein and Slovic 1971; Grether and Plott 1979; Berg, Dickhaut and Rietz 2010). If success payoffs of delegation lotteries are larger than the ones of control lotteries, subjects will give, according to this explanation, a higher certainty equivalent to delegation lotteries. We need then to check if delegation lotteries are considered as high-amount lotteries, i.e., if $P_{\mathcal{A}}$ is larger than $P_{\mathcal{P}} - C(E)$. We find that control lotteries have a smaller success payoff in 56.1% of the cases (59.3% in France, and 52.7% in Japan), a larger one in 43.7% of the cases (40.6% in France, and 47% in Japan) and the payoffs are equal in 0.3% of the cases (2 cases in France, and 7 in Japan). Moreover, our findings indicate that the control and delegation lotteries have similar probabilities. The average success probabilities are, respectively, 63.6% (65.6% in France and 61.4% in Japan) and 59.8% (63.5% in France and 55.8% in

Japan). In 43.6% of cases (43.9% in France and 43.2% in Japan), the control lottery has a higher probability of success than the delegation lottery. On the contrary, delegation is the high-probability lottery in 36.7% of cases (39.7% in France and 33.4% in Japan). The two lotteries have the same probability of success in 19.7% of cases (16.3% in France and 23.4% in Japan). Taken together, these results suggest that preference reversals are not behind the intrinsic value of decision rights observed in our experiment.

Finally, corner minimum effort requirements, i.e., $\underline{e} = 1$ and $\underline{e} = 100$, could in principle undermine the elicitation of the principals' point of indifference. Nonetheless, we observe low percentages of corner solutions for $\underline{e}$. In our complete sample, principals selected $\underline{e} = 1$ in 6.2% of cases. In total, 75.2% of subjects have never chosen $\underline{e} = 1$ (84.5% in France, and 65.2% in Japan), 8.8% of them have chosen $\underline{e} = 1$ only once (7% in France, and 10.6% in Japan), and 8% twice (7% in France, and 9.1% in Japan). It turns out that $\underline{e} = 100$ is chosen in 5.29% of cases, and 81.4% of subjects have never chosen $\underline{e} = 100$ (83.8% in France, and 78.8% in Japan), 6.9% of them have chosen $\underline{e} = 100$ only once (4.2% in France, and 9.9% in Japan), and 5.5% twice (7% in France, and 3.8% in Japan). These findings, both for $\underline{e} = 1$ and $\underline{e} = 100$, are similar for all the treatments.[42] Given its low frequency, and in accordance with the additional control experiment ran by Bartling et al. (2014) to address this issue, we conclude that these choices do not pose a problem to the elicitation of the principals' point of indifference.

[42]The higher frequency is for treatment 1 and $\underline{e} = 1$ in Japan, in which 46.5% of subjects have never chosen $\underline{e} = 1$, 18.6% of them have chosen $\underline{e} = 1$ once, and 11.6% twice.

5.G Instructions

This Appendix contains the English instructions of the experiment that were handed out to subjects in the position of principals in the three treatments, and from which the French and Japanese instructions were translated and back translated. The French and Japanese instructions, as well as the agent's instructions of Part 1 (the remaining instructions are common to principals and agents) are available from the authors upon request.

To be self-contained, we exclude repetitions of the instructions between treatments. The red-colored sentences are specific to different treatments, with [T1] indicating sentences for treatment 1, [T2] sentences for treatment 2, and [T3] sentences for treatment 3. The black-colored sentences are common to all instructions, based on treatment 1 and with the exception (not highlighted) that "Participant B" is substituted for "the bot" in treatment 2 and sometimes in treatment 3 (only in sentences where "Participant B" in the instructions refer to the one that makes decisions or the potential holder of the decision right). This appendix is organized as follows:

A. Instructions for Part 1 (Principals)

B. Instructions for Part 2 (All subjects)

C. Instructions for Illusion of Control Task (All subjects)

D. Instructions for Loss Aversion Task (All subjects)

E. Instructions for Cognitive Ability Task (All subjects)

F. Supplementary Cost Sheet (All subjects)

A. Instructions for Part 1 (Principals)

Instructions for Participant A

Welcome to this experiment.

Please carefully read the following instructions. These will provide you with all the information needed to participate in this experiment. If you don't understand something, don't hesitate to raise your hand. We will come and answer your question where you are seated.

You will receive an initial endowment of **5 euros** at the start of the experiment. You can earn an additional monetary amount during the experiment by earning **points**. The number of points you will earn depends on both your decisions and those of other participants.

All the points you earn during the course of this experiment will be converted to euros at the end of the experiment. The following exchange rate will be applied:

100 points = 2.50 euros

At the end of the experiment, you will receive the money you earned during the experiment as well as the initial sum of 5 euros.

Please note that **all communication is strictly forbidden during the entirety of the experiment**. We also want to emphasize that you must only use the computer functions that are related to the experiment. We remain at your disposal to answer any questions you might have.

This experiment is composed of **4 parts**:

1. [T1 and T3] The first part of the experiment is composed of 10 rounds. For each of these 10 rounds, you will be randomly paired with a Participant B. You will be able to implement a project with the Participant B who is randomly paired with you in each round. A detailed explanation of the first part of the experiment is found in the following pages.
1. [T2] The first part of the experiment is composed of 10 rounds. For each of these 10 rounds, you will interact with a ``bot''. You will be able to implement a project with the bot in each round. A detailed explanation of the first part of the experiment is found in the following pages.

2. In the second part of the experiment, you will be presented with 20 different decisions between a fixed and an unfixed amount. You will receive detailed instructions on the second part of the experiment once the first part is concluded.

3. The third part of the experiment is very short and you will receive detailed instructions once the second part is concluded.
4. In the fourth part of the experiment, we will ask you to answer a series of questions.

General Instructions for the First Part of the Experiment

[T1 and T3] There are two types of participants in the first part of the experiment: Participant A and Participant B. **You are Participant A**.

There are 10 rounds. You will be paired with a different Participant B in each round. A **project** can be implemented in each round. If the project is a success, Participant A and B will receive positive payments. A successful implementation of the project will lead to a positive payment for participants A and B.

[T2] There are 10 rounds. You will be paired with a **bot** for each round. A **project** can be implemented in each round. If the project is a success, you will receive a positive payment.

The decision right

In each round, either you or Participant B has the decision right. The participant with the decision right can make two decisions:

1. **Which project – A or B – will be implemented?**
   Participant A receives a greater share of the project payment in Project A and Participant B receives a greater share of the project payment in Project B (It is possible that Participant A and Participant B receive the same share in certain rounds).

2. **What is the probability the project will be successful?**
   The determination of the probability of success is connected to the costs paid by the participant who has the decision right. The higher the probability of success, the higher the costs.

[T3] Please note that if the bot has the decision right, the bot makes the decisions **on behalf of Participant B**. To put it another way, Participant B will not make their decisions themselves. This means that the payment of the project and the costs linked to the choice of the probability of success (a choice made by the bot) are automatically assigned to Participant B.

[T2 and T3] The manner in which the bot makes decisions is described later in the instructions.

Payment of the project

The payments that result from the implementation of the project vary from one round to another. You will be informed of the payments at the start of each round.

Example: The payments for one project for one round. In the case that Project A is successful, you receive 200 points and Participant B receives 150 points. In the case that Project B is successful, Participant B receives 200 points and you receive 150 points. In the event that the project fails, both participants receive 100 points each.

		Your Payment	Payment for Participant B
In the case of success	**Project A**	200	150
	Project B	150	200
In the case of failure		100	100

The probability of success

If you have the decision right, then you can determine the probability of success for the chosen project, either A or B.

How is the probability of success determined?

The probability of success is a number between 0 and 100 that can be chosen freely.

$$\mathbf{0 \leq \text{probability of success} \leq 100}$$

A probability of success of 0 means that the project will never be successful. A probability of success of 100 indicates that the success of the project is guaranteed. A value of 50 indicates that a project has a 50% chance of success.

The cost of the choice of the probability of success

The higher the probability of success you choose, the higher the cost. Two information sheets (one blue and one yellow) are at your desk: they both provide a table and a graph outlining the cost schedule for the different probabilities of success. Each round, you will be informed whether the cost schedule from the blue or yellow sheet will be applied. You can also always have the computer show you the costs on the monitor while choosing the probability of success.

The success of the project

At the end of the experiment, one of the 10 rounds will be randomly selected by the computer. The choices made by you and Participant B in this round will determine your payments for the first part of the experiment.

The success or failure of the project chosen by the participant with the decision right for the randomly selected round will be determined in the following manner.
[T2 and T3] The success or failure of the project chosen by the one with the decision right (either you or the bot) for the randomly selected round will be determined in the following manner.
A number between 1 and 100 will be drawn; all numbers between 1 and 100 have an equal chance of being drawn. The number that is drawn will then be compared to the probability of success that was chosen by the participant with the decision right.

If the number drawn is **smaller than or equal** to the probability of success that was chosen, the project is a success. If the number drawn **is larger**, the project is not a success. **The greater the probability of success that you have chosen, the greater the chance that the**

number drawn will be smaller than your chosen probability. To put it another way, there is a greater chance your project will be successful.

Examples:

1. **Example 1**: You have chosen a **probability of success of 15**, that is to say 15%.

   This means:
   - If the number drawn at random is between 1 and 15 (= 15 chances out of 100), the project is successful.
   - If the number drawn is larger than 15 (= 85 chances out of 100), the project is not a success.

2. **Example 2**: You have chosen **a probability of success of 80**, that is to say 80%.

   This means:
   - If the number drawn at random is between 1 and 80 (= 80 chances out of 100), the project is successful.
   - If the number drawn is larger than 80 (= 20 chances out of 100), the project is not a success.

- **Suppose that the number chosen at random is 93.**
  In this case, the project is not a success in either example (the randomly drawn number is larger than the chosen probability of success in both examples).

- **Suppose that the number chosen at random is 54.**
  In this case, the project in Example 1 would not have been successful (the randomly drawn number is larger than 15) but the project in Example 2 would have been a success (the randomly drawn number is less than 80).

- **Suppose that the number chosen at random is 3.**
  In this case, the project would have been a success in both examples (the randomly drawn number is lower than the chosen probability of success in both examples).

The income

The incomes for Participant A and Participant B are made up of two elements:

- The payment from the chosen project in the event the project is successful. In the case the project fails, the two participants receive a lower payment that is independent of the project chosen.
- The costs linked to the chosen probability of success are deducted from the payment of the participant who has the decision right.

This results in the following four possibilities for you:

1. You have the **decision right** and the project is **successful**:

Income = Payment from the project that you chose minus the costs linked to the choice of the probability of success

2. You have the **decision right** and the project is a **not a success**:

Income = Payment in case of failure minus the costs linked to the choice of the probability of success

3. You **do not have the decision right** and the project is **successful:**

Income = Payment from the project chosen by Participant B

[T2] Income = Payment from the project chosen by the bot

[T3] Income = Payment from the project that the bot chose on behalf of Participant B

4. You **do not have the decision right** and the project is a **not a success:**

Income = Payment in case of failure

[T2] Please note that the bot's payments are hypothetical. **Nobody in the room will receive the points earned by the bot during the first part of the experiment.**

[T3] Please note that **Participant B makes no decisions.** Thus, they have no influence on your income. But the decisions that the bot makes in their place as well as your decisions will affect the income of Participant B.

Detailed Procedure of One Round on the Computer

[T1] 1st Stage: Participant B's decision

In each round, you as participant A first have the decision right. You can also opt to delegate the decision right to Participant B. Before deciding if you want to delegate the decision to Participant B, Participant B must make a definite choice of a project and a probability of success in the event that you delegate the decision right.

If you end up delegating the decision right to Participant B, then the decisions participant B makes in the first stage will be realized.

You will not yet learn which decisions participant B makes in the first stage.

[T2] 1st Stage: Bot's decision

In each round, you as participant A first have the decision right. You can also opt to delegate the decision right to the bot. Before deciding if you want to delegate the decision to the bot, the bot must make a definite choice of a project and a probability of success in the event that you delegate the decision right.

How does the bot make its decisions?

1. The bot always chooses the project that earns itself the most points.
1. The bot chooses a probability of success between 0 and 100 **at random**. There is thus a 1/101 chance that the bot picks a probability of success equal to 0; a 1/101 chance that the bot picks a probability of success equal to 1; etc.; and a 1/101 chance that the bot picks a probability of success equal to 100.

If you end up delegating the decision right to the bot, then the decisions the bot makes in the first stage will be realized.
You will not yet learn which decisions the bot makes in the first stage.

[T3] 1st Stage: Bot's decision on behalf of Participant B

In each round, you as participant A first have the decision right. You can also opt to delegate the decision right to the bot. Before deciding if you want to delegate the decision to the bot, the bot must make a definite choice of a project and a probability of success (on behalf of Participant B) in the event that you delegate the decision right.

How does the bot make its decisions?

1. The bot always chooses the **project that earns Participant B the most points**.
2. The bot chooses a probability of success between 0 and 100 **at random**. There is thus a 1/101 chance that the bot picks a probability of success equal to 0; a 1/101 chance that

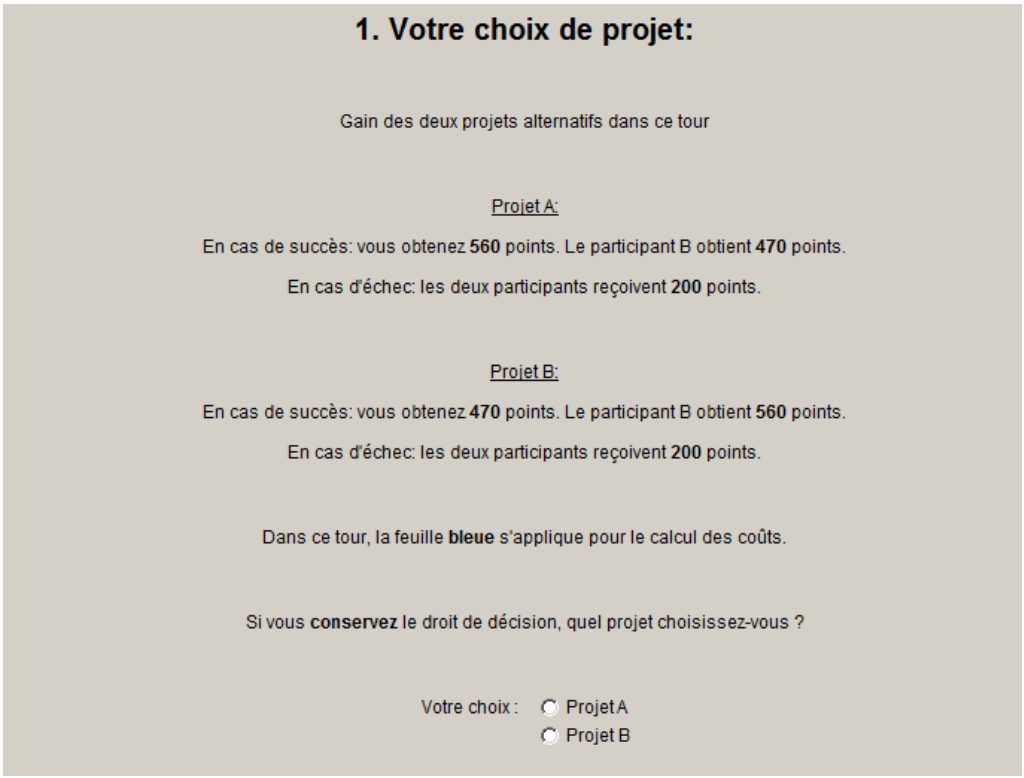
the bot picks a probability of success equal to 1; etc.; and a 1/101 chance that the bot picks a probability of success equal to 100.

If you end up delegating the decision right to the bot, then the decisions the bot makes in the first stage will be realized.

You will not yet learn which decisions the bot makes in the first stage.

2nd Stage: Choice of project

At this stage of the experiment, you have not yet made the final decision on whether or not you will delegate the decision right. For this reason, you must select the project that you would like to implement in case you opt to keep the decision right. The choice of project will be made on this type of screen:

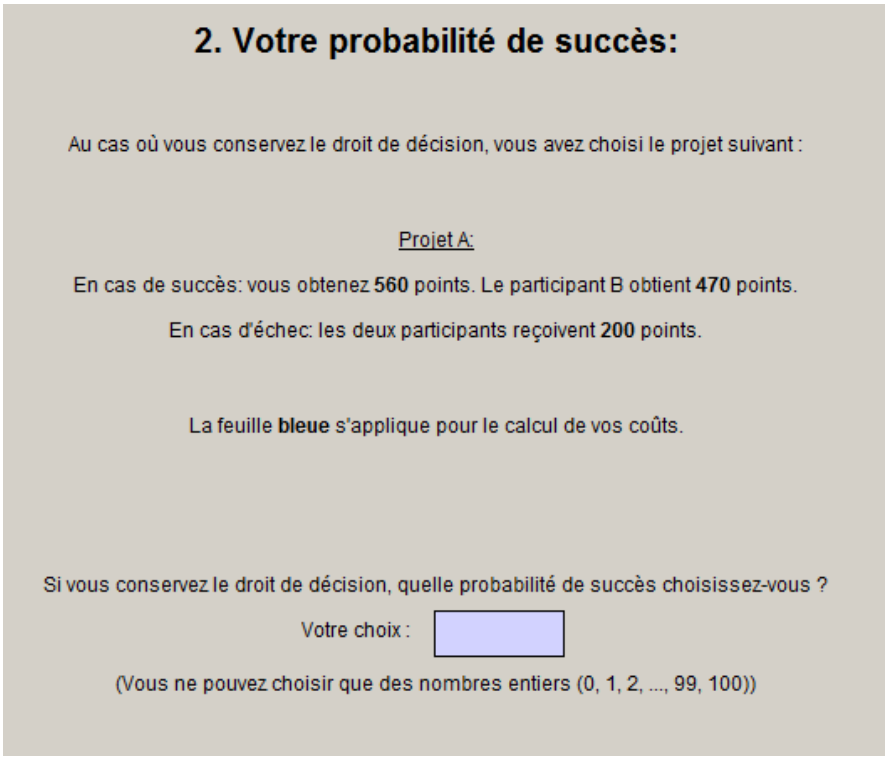


After having chosen a project, please click on the "OK" button.

3rd Stage: Choice of probability of success

When selecting the probability of success, you still have not made a definite choice of whether or not to delegate the decision right. After having chosen a project, you must select the probability of success for this choice in case you keep the decision right. The cost of the probability of success will only be applied if you ultimately keep the decision right.

You make your choice of the probability of success on this type of screen:



After having selected the probability of success, click on the button "Display costs". This will then show you the exact costs of the probability of success that you chose. You can modify your probability of success if you wish. By clicking on "Confirm", you make your definitive selection.

4th Stage: Who has the decision right?

You can decide in each round – after participant B has made his decisions – whether you would like to delegate the decision right to participant B or if he would like to retain this for yourself. In this case, you do not make the decision directly, but by **determining a minimum requirement**:

[T3] You can decide in each round – after the bot has made his decisions on behalf of Participant B – whether you would like to delegate the decision right to the bot or if he would like to retain this for yourself. In this case, you do not make the decision directly, but by **determining a minimum requirement**:

In each round, you determine the minimum probability of success that Participant B must have chosen in order for you to be willing to delegate the decision right to them. You can choose any minimum requirement between 1 and 100.

Participant B has already chosen their probability of success at the moment when you determine a minimum requirement. Thus there is no possibility you will influence the choice made by Participant B.

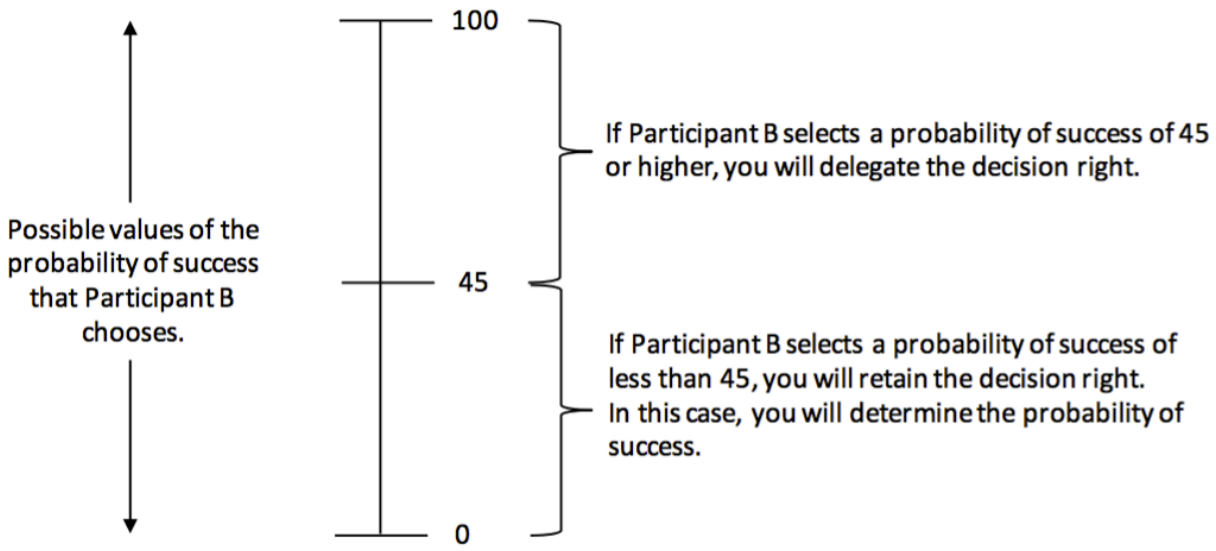
Please note that you do not know the probability of success chosen by Participant B when you determine your minimum requirement.

[T3] Please also note that Participant B will make no decisions.

If the probability of success chosen by Participant B is equal to or higher than the minimum requirement you have determine, you will delegate the decision right. If the probability of success chosen by Participant B is lower than the minimum requirement you determine, you will keep the decision right.

The graph seen below will clarify the link between the minimum requirement you have determine, the probability of success chosen by Participant B, and the question of who will have the decision right.

If, for example, you chose **a minimum requirement of 45,** this means that you wish to delegate the decision right to Participant B if they have selected a probability of success of 45 or more.



When considering what minimum requirement to determine, you should ask the following question:

- Do I want to delegate the decision right if Participant B selects a probability of success of 1? If the answer is no, you should ask the question:
- Do I want to delegate the decision right if Participant B selects a probability of success of 2? If the answer is no, you should ask the question:

- Do I want to delegate the decision right if Participant B selects a probability of success of 3? And so on.

Do this until you reach a level of probability of success chosen by Participant B above which you would delegate the decision right. This level should be your minimum requirement.

- In the above example, the value is 45. This means that you would just be willing to delegate the decision right if Participant B chooses a probability of success of 45 but that you would prefer retaining this right at all values of 44 or less.

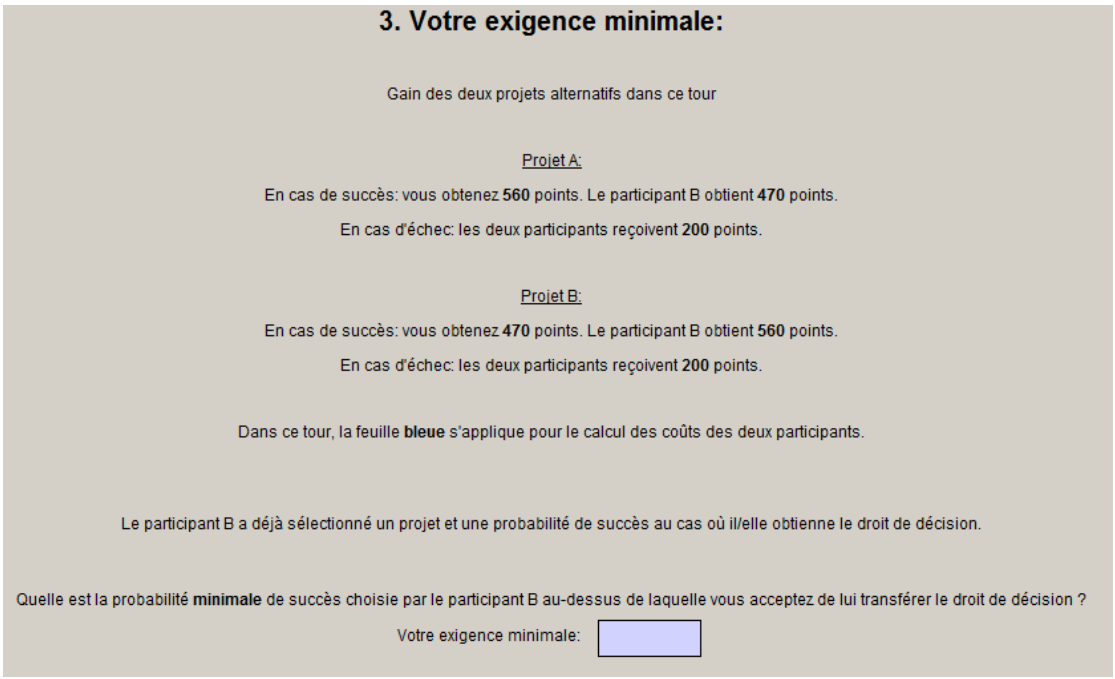
Other examples:

1. You select **a minimum requirement of 78**.
   This means:
   - If during Stage 1, Participant B selects a probability of success between 0 and 77, you do not delegate the decision right.
   - If during Stage 1, Participant B selects a probability of success between 78 and 100, you delegate the decision right to them.

2. You select **a minimum requirement of 4**.
   This means:
   - If during Stage 1, Participant B selects a probability of success between 0 and 3, you do not delegate the decision right.
   - If during Stage 1, Participant B selects a probability of success between 4 and 100, you delegate the decision right to them.

You make your decision on the minimum requirement for participant B on a screen like the one shown below:

The upper part of the screen informs you of the payments in the two project alternatives as well as the payment in case of lack of success in the round in question. You will also be informed whether the cost schedule on the blue or yellow sheet will be applied for the round in question. You can indicate your choice of a minimum requirement for delegating the decision right in the lower part of the screen. Here is an example:



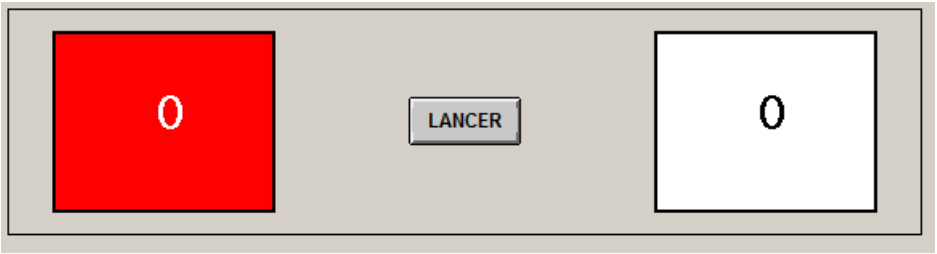
After having indicated your minimum requirement, please click on the "OK" button. A new round will then begin.

How the success of the project is determined

At the end of the experiment, the computer will randomly select one of the 10 rounds and the payments for you and Participant B for this part of the experiment will be determined on the basis of the decisions made by you and Participant B **in that round**. As you will not know which round will be randomly selected by the computer, you should make careful decisions every round.

a) The computer will first randomly determine which round will be selected for payment.
b) The computer then assesses whether the Participant B you had been randomly paired with for that round chose a probability of success that was at least equal to your minimum requirement.
   - If the minimal requirement is lower than or equal to the probability of success chosen by Participant B, you will delegate the decision right.
   - If the minimal requirement is greater than the probability of success chosen by Participant B, you will retain the decision right.

If you keep the decision right, the success of the project you chose during the randomly-selected round is determined by a pair of **electronic dice** that will randomly pick a number between 1 and 100. This number is then compared to the probability of success you have chosen.
More precisely, the participant who holds the decision right will see the following on their screen:



The number on the red background represents the tens column and the number on the white background represents the ones column.

Please then click on the "THROW" button. You will see numbers on the dice **change quickly in a random manner**. You can stop the numbers by clicking on the button "STOP".
As you will see, the numbers change too quickly to be able to choose which numbers to stop on.

After having clicked on "STOP", the two numbers that appear on the screen will give you a number between 1 and 100 (two zeroes represent 100).
If the number is less than or equal to the probability of success chosen by the participant who had the decision right, the project is then a success. On the contrary, if the number is greater than the probability of success chosen by the participant who had the decision right, the project is then a failure.

[T2 and T3] If you did not have the decision right, a number between 1 and 100 will be chosen at random by the computer. If this number is less than or equal to the probability of success chosen by the bot at random, the project is then a success. On the contrary, if the number is greater than the probability of success chosen by the bot, the project is then a failure.

Summary of the First Part of the Experiment

These are the stages of one round:

1st Stage: Participant B chooses a project and a probability of success in the event that you delegate the decision right to them.

2nd Stage: You choose a project in the event you keep the decision right.

3rd Stage: You choose a probability of success in the event you keep the decision right.

4rd Stage: You choose the minimum requirement for the probability of success for the project that Participant B must choose in order for you to agree to delegate the decision right to them.

At the end of the experiment, one of the 10 rounds will be drawn at random. The decisions that you and Participant B made during that round will determine the monetary income for the first part of the experiment. This will be added to your initial payment of 5 euros and the payments you obtain in the rest of the experiment.

Do you have any questions regarding this experiment? Please raise you hand if you have one. We will come answer you where you are seated.

You will find questions to test your understanding of the experiment on the following pages.

Comprehension Questions

Please answer the following comprehension questions. Please signal the experiment supervisor if you have any questions.

1. Consider the case where you have selected a minimum requirement of 85.

a) If Participant B selected a probability of success of 80, who has the decision right in this round? ……

b) If Participant B selected a probability of success of 90, who has the decision right in this round? ……

2. Consider the case where you have selected a minimum requirement of 55.

a) If Participant B selected a probability of success of 50, who has the decision right in this round? ……

b) If Participant B selected a probability of success of 60, who has the decision right in this round? ……

3. Consider the case where Participant B chose a probability of success of 3.

a) If you chose a minimum requirement of 1, who has the decision right in this round? ……

b) What is the probability that the project will be successful? ……

c) If you chose a minimum requirement of 4, who has the decision right in this round? ……

d) What is the probability that the project will be a successful? ……

4. Consider the case where Participant B chose a probability of success of 90.

a) If you chose a minimum requirement of 85, who has the decision right in this round? ……

b) What is the probability that the project will be successful? ……

c) If you chose a minimum requirement of 95, who has the decision right in this round? ……

d) What is the probability that the project will be successful? ……

5. Consider the case where you keep the decision right and you have chosen a probability of success of 54. The cost schedule from the yellow information sheet applies in this round. Assume further that you obtain 8 as your number on the red background and 2 as your number on the white background.

a) What are your costs? ……

b) Will the project be successful? ……

The following payments are applied for this project:

		Your Payment	Payment for Participant B
In case of success	**Project A**	200	150
	Project B	150	200
In case of failure		100	100

Consider the case that you have chosen Project A.

c) What will be your payment? ……

d) What will be the payment for Participant B?

Now consider the case where you have chosen a probability of success of 24. Also, you now have the number 1 on the red background and the number 5 on the white background. The cost schedule from the yellow paper is applied for this round.
You have again chosen Project A.

e) What are your costs? ……

f) Will the project be successful? ……

g) What will be your payment? ……

h) What will be the payment for Participant B? ……

6. Consider the case where you have delegated the decision right. Participant B selected Project B and chose a probability of success of 48. The cost schedule from the blue paper is applied for this round.

The following payments are applicable to the project:

		Your Payment	Payment for Participant B
In case of success	**Project A**	180	150
	Project B	150	180
In case of failure		100	100

Consider the case where Participant B obtains the number 5 on the red background and the number 7 on the white background.

a) Will the project be successful? ……

b) What will your payment be? ……

c) What will be the payment of Participant B? ……

Consider the case where Participant B obtains the number 3 on the red background and the number 9 on the white background.

d) Will the project be successful? ……

e) What will your payment be? ……

f) What will be the payment of Participant B? ……

Comprehension Questions: Answers

Please answer the following comprehension questions. Please signal the manager of the experiment if you have any questions.

1. Consider the case where you have selected a minimum requirement of 85.

   a) If Participant B selected a probability of success of 80, who has the decision right in this round? **You**

   b) If Participant B selected a probability of success of 90, who has the decision right in this round? **Participant B**

2. Consider the case where you have selected a minimum requirement of 55.

   a) If Participant B selected a probability of success of 50, who has the decision right in this round? **You**

   b) If Participant B selected a probability of success of 60, who has the decision right in this round? **Participant B**

3. Consider the case where Participant B chose a probability of success of 3.

   a) If you chose a minimum requirement of 1, who has the decision right in this round? **Participant B**

   b) What is the probability that the project will be successful? **3%**

   c) If you chose a minimum requirement of 4, who has the decision right in this round? **You**

   d) What is the probability that the project will be a success? **According to your choice**

4. Consider the case where Participant B chose a probability of success of 90.

   a) If you chose a minimum requirement of 85, who has the decision right in this round? **Participant B**

   b) What is the probability that the project will be successful? **90%**

   c) If you chose a minimum requirement of 95, who has the decision right in this round? **You**

d) What is the probability that the project will be successful? **According to your choice**

5. Consider the case where you keep the decision right and you have chosen a probability of success of 54. The cost schedule from the yellow information sheet applies in this round. Assume further that you obtain 8 as your number on the red background and 2 as your number on the white background.

a) What are your costs? **29.2 points**

b) Will the project be successful? **No**

The following payments are applied for this project:

		Your Payment	Payment for Participant B
In case of success	**Project A**	200	150
	Project B	150	200
In case of failure		100	100

Consider the case that you have chosen Project A.

c) What will be your payment? **100 – 29.2 = 70.8**

d) What will be the payment for Participant B? **100**

Now consider the case where you have chosen a probability of success of 24. Also, you now have the number 1 on the red background and the number 5 on the white background. The cost schedule from the yellow paper is applied for this round.
You have again chosen Project A.

e) What are your costs? **5.8**

f) Will the project be successful? **Yes**

g) What will be your payment? **200 – 5.8 = 194.2**

h) What will be the payment for Participant B? **150**

6. Consider the case where you have delegated the decision right. Participant B selected Project B and chose a probability of success of 48. The cost schedule from the blue paper is applied for this round.

The following payments are applicable to the project:

		Your Payment	Payment for Participant B
In case of success	**Project A**	180	150
	Project B	150	180
In case of failure		100	100

Consider the case where Participant B obtains the number 5 on the red background and the number 7 on the white background.

a) Will the project be successful? **No**

b) What will your payment be? **100**

c) What will be the payment of Participant B? **100 – 46.1 = 53.9**

Consider the case where Participant B obtains the number 3 on the red background and the number 9 on the white background.

d) Will the project be successful? **Yes**

e) What will your payment be? **150**

f) What will be the payment of Participant B? **180 – 46.1 = 133.9**

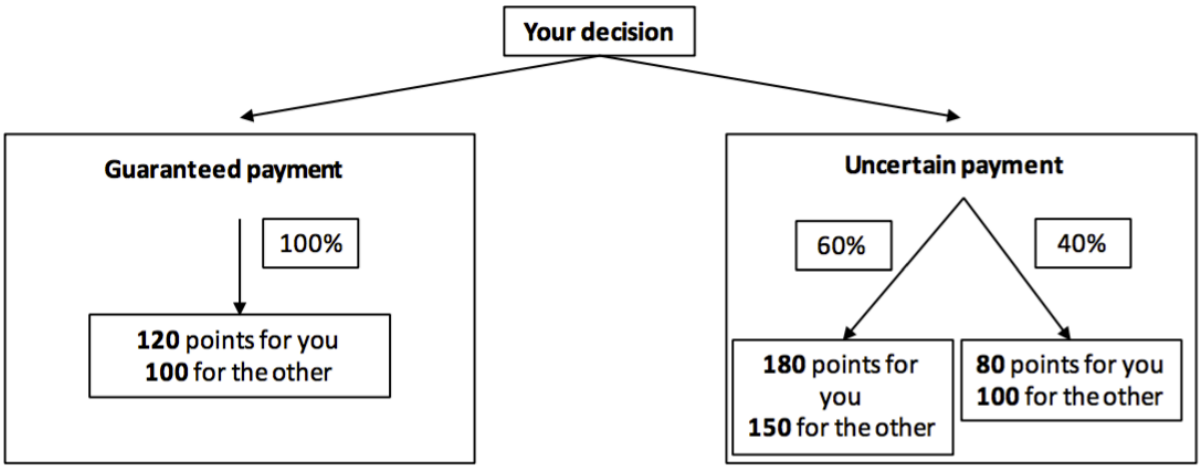
B. Instructions for Part 2 (All subjects)

Second Part of the Experiment

The second part of the experiment is made up of 20 rounds. You will be randomly paired with another participant for every round. The exchange rate of 2.50 euros for 100 points is still applicable.

In each round, you must choose between a guaranteed payment and an uncertain payment. Your choice will also affect the payment of the other participant with whom you have been randomly paired.

Here's an example:



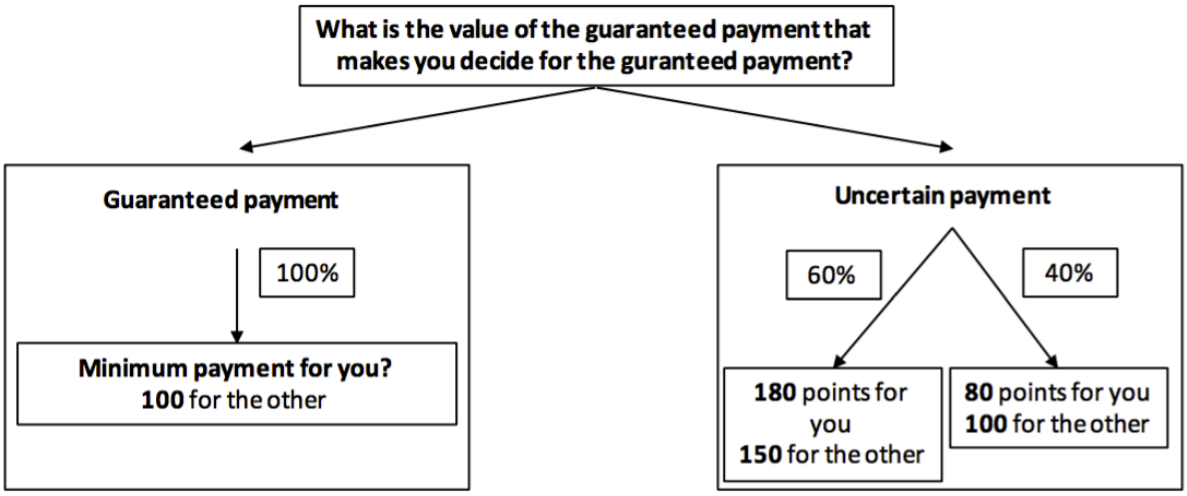
If, in the example above, you decide for the guaranteed payment, you will receive 120 points and the participant you have been randomly associated will receive 100 points.

If you opt for the uncertain payment, there is a 60% probability that you will receive 180 points and the other participant will receive 150 points. There is a 40% probability you will receive 80 points and the other participant will receive 100 points.

In each of the 20 rounds, you choose between a guaranteed payment and an uncertain payment. The amount of the payments and the probabilities change each round.

How do you choose between guaranteed payment and uncertain payment in each round?

When you make your choice between the guaranteed payment and the uncertain payment, you don't yet know the amount of the guaranteed payment. Thus, you can't directly choose between the guaranteed payment and the uncertain payment. Instead, you must indicate the minimum payment for which you prefer to choose the guaranteed payment than the uncertain payment.



In each round, you will be informed of the guaranteed payment for the other participant, the uncertain payment available to you and the other participant, and the probabilities associated with the uncertain payment in each round.

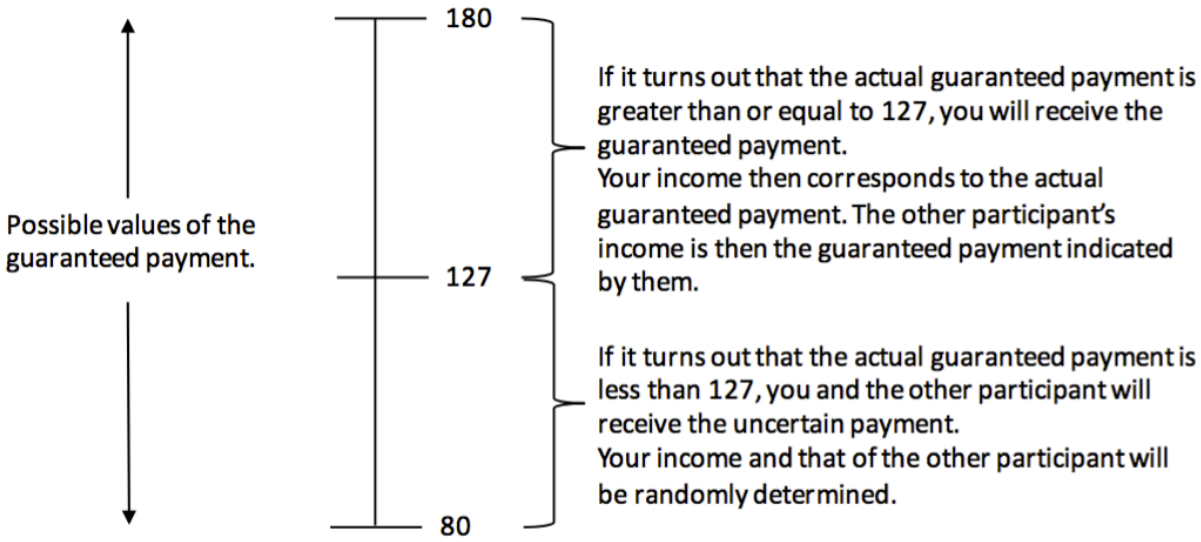
After having indicated the **minimum payment** that would make you choose the guaranteed payment in a round, you will be informed of your **actual guaranteed payment** for that round. The choice between the guaranteed payment and the uncertain payment is made in the following manner:

- If the guaranteed payment is less than the minimum payment that you have chosen, the uncertain payment determines your income and that of the other participant.
- If the guaranteed payment is equal to or greater than the minimum payment you have chosen, you will receive the actual guaranteed payment and the other participant will receive the guaranteed payment shown on the screen (100 points in the above example).

The possible values of your guaranteed payment lie between both of your uncertain payments (in the above example, between 80 and 180 points). Every amount in this range (80, 81,..., 180) is possible and has the same probability of occurring. The minimum guaranteed payment you choose can be any integer value between the two possibilities for your uncertain payment.

The following graph clarifies the relationship between the minimum payment you choose, the amount of the actual guaranteed payment, and your choice between the guaranteed payment and the uncertain payment.

If, for example, you choose a **minimum payment of 127**, this means that you prefer any guaranteed payment between 127 and 180 points to the uncertain payment.
You will only learn the exact value of your **actual guaranteed payment** after you have chosen your minimum payment.



When you consider your minimum payment, then you should (assuming the numbers from the example above) ask the following questions:

- Would I prefer a guaranteed payment of 180 points to the uncertain payment? If yes, you then should ask:
- Would I prefer a guaranteed payment of 179 points to the uncertain payment? If yes, you then should ask:
- Would I prefer a guaranteed payment of 178 points to the uncertain payment? Etc.

Continue until you reach a value for the guaranteed payment where you just prefer the guaranteed payment to the uncertain payment. You should then choose this value as your minimum payment.

In the above example, this value is 127. This means you have a slight preference for the guaranteed payment of 127 over the uncertain payment but you prefer the uncertain payment over the guaranteed payment of 126 (or any guaranteed payment less than 126).

The income:

If the actual guaranteed payment is at least as high as the minimum payment you selected:

You will receive the actual guaranteed payment.
The other participant will receive the guaranteed payment assigned for them.

If the actual guaranteed payment is less than the minimum payment you selected:

One of the two possible uncertain payments will be selected randomly according to the indicated probabilities.

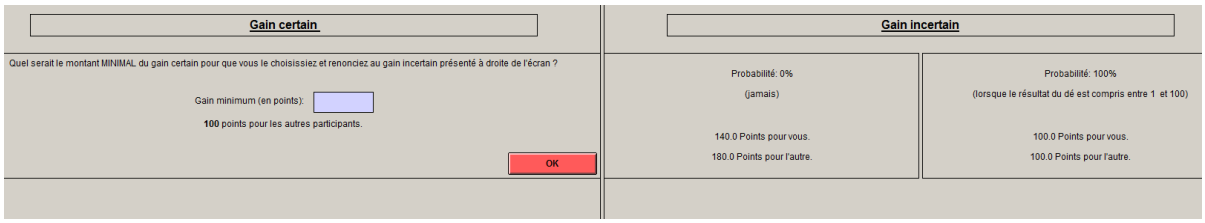
At the end of the experiment, the computer will pick 2 of the 20 rounds at random.

For each of the randomly selected rounds, the minimum payment you have chosen will be compared to the actual guaranteed payment. If the actual guaranteed payment is equal to or greater than the minimum payment you have chosen, you will receive the guaranteed payment. If the actual guaranteed payment is less than the minimum payment you have chosen, a random draw will determine which of the uncertain payments you and the other participant each receive.

As you don't know which rounds will be randomly drawn by the computer, it is in your interest to make careful decisions every round.

Procedure on the Computer

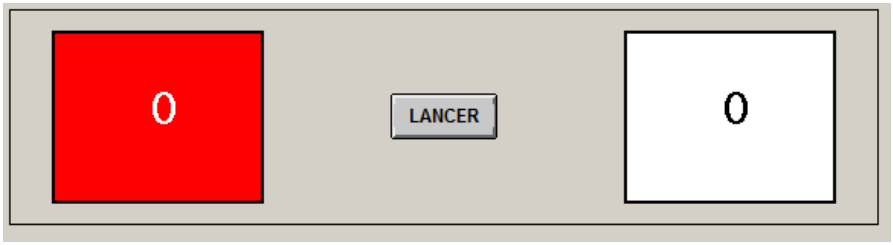
For each round, you will choose the guaranteed payment that you must receive as a minimum in order to make you prefer the guaranteed payment over the uncertain payment. This decision will be made each round on this type of screen:



On the right of the screen you can see the possible values for your uncertain payments and the uncertain payments for the other participant who has been paired with you in a random manner. You will also see the probability of receiving each of the possible payments. This information changes for each of the 20 rounds.

You indicate the **minimum payment** on the left of the screen. This minimum payment indicates what value the guaranteed payment must be for you to prefer the guaranteed payment to the uncertain payment. Once you have entered your choice, please click on the "OK" button. You can modify the number you enter up until the point you click on the "OK" button.

At the end of the experiment, two rounds will be drawn randomly. If the actual guaranteed payment is less than the minimum payment you have chosen in one of these two rounds, a number between 1 and 100 will be selected randomly by a pair of **electronic dice**. More precisely, you will see the following image on your screen:



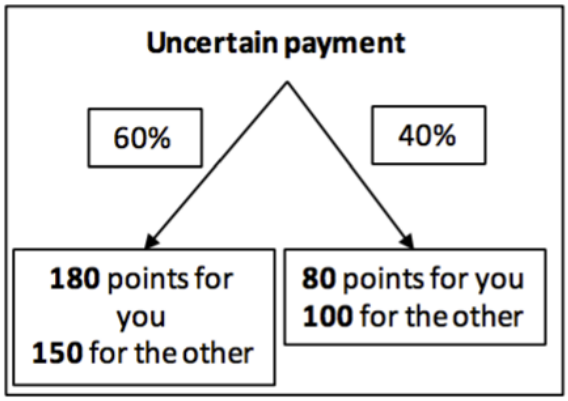
The number on the red background represents tens and the number on the white background represents ones (two zeros indicate 100).

Please then click on the "THROW" button. You will see numbers on the dice **change quickly in a random manner**. You can stop the numbers by clicking on the "STOP" button.
As you will see, the numbers change too quickly to be able to choose which numbers to stop on.

After having clicked on "STOP", the two numbers that appear on the screen will give you a number between 1 and 100 (two zeros represent 100).

The number on the screen will then determine which of the uncertain payments you and the other participant (who you have been randomly paired with for this part of the experiment) will receive.

For example, consider the scenario with the following uncertain payments:



If you ended up with an uncertain payment rather than a guaranteed payment, a number between 1 and 100 will be selected at random using the electronic dice.

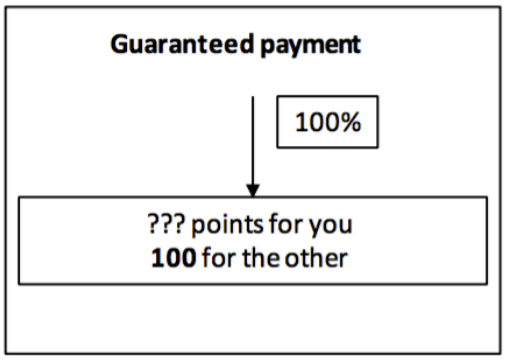
- If the number is between 1 and 60, then you will receive 180 points and the other participant will receive 150 points.
- If the number is between 61 and 100, then you will receive 80 points and the other participant will receive 100 points.

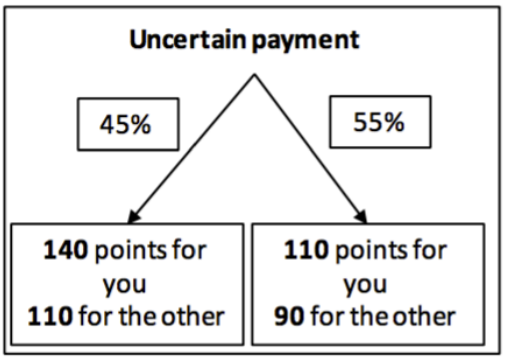
Do you have any questions regarding the second part of the experiment? If so, please raise your hand. We will come to your seat to give you an answer.

You will find questions to test your understanding of the experiment on the following pages.

Comprehension Questions

Consider a scenario with the following sums and probabilities:





1. **In the case that you chose a minimum payment of 120:**
   (b) If the actual guaranteed payment is 128:
   What will your payment be for this round? .....
   What will the payment be for the other participant in this round? .....

   (b) If the actual guaranteed payment is 117:
   What will your payment be for this round? .....
   What will the payment be for the other participant in this round? .....

2. **In the case that you chose a minimum payment of 135:**
   (a) In the case the actual guaranteed payment is 128:
   What will your payment be for this round? .....
   What will the payment be for the other participant in this round? .....

   (b) In the case the actual guaranteed payment is 113:
   What will your payment be for this round? .....
   What will the payment be for the other participant in this round?

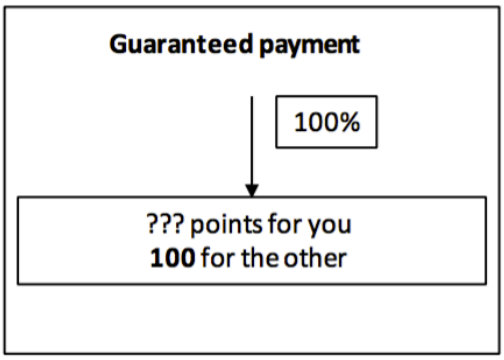
3. **In the case that you chose a minimum payment of 115:**
   (a) In the case the actual guaranteed payment is 128:
   What will your payment be for this round? .....
   What will the payment be for the other participant in this round?.....

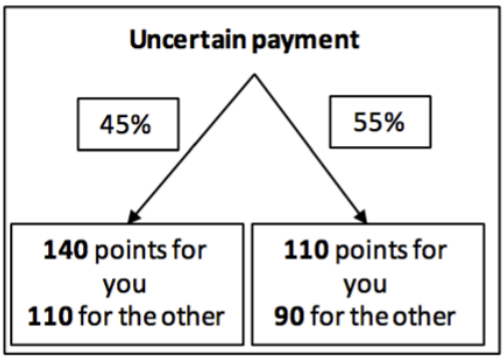
   (b) In the case the actual guaranteed payment is 135:
   What will your payment be for this round? .....
   What will the payment be for the other participant in this round?.....

If you have a question, please raise your hand. We will come answer you where you are seated.

Comprehension Questions: Answers

Consider a scenario with the following sums and probabilities:





A. In the case that you chose a minimum payment of 120:

(b) In the case the actual guaranteed payment is 128:
What will your payment be for this round? **128**
What will the payment be for the other participant in this round? **100**

(b) In the case the actual guaranteed payment is 117:
What will your payment be for this round? **140 or 110**
What will the payment be for the other participant in this round? **110 or 90**

B. In the case that you chose a minimum payment of 135:

(a) In the case the actual guaranteed payment is 128:
What will your payment be for this round? **140 or 110**
What will the payment be for the other participant in this round? **110 or 90**

(b) In the case the actual guaranteed payment is 113:
What will your payment be for this round? **140 or 110**
What will the payment be for the other participant in this round? **110 or 90**

C. In the case that you chose a minimum payment of 115:

(a) In the case the actual guaranteed payment is 128:
What will your payment be for this round? **128**
What will the payment be for the other participant in this round? **100**.

(b) In the case the actual guaranteed payment is 135:
What will your payment be for this round? **135**
What will the payment be for the other participant in this round? **100**

If you have a question, please raise your hand. We will come answer you where you are seated.

C. Instructions for Illusion of Control Task (All subjects)

Additional Information

The computer will select a round at random to use as the basis to calculate your payments for the first part of the experiment. If you have the right to decide for the chosen round, you will be able to stop the electronic dice yourself.

We would like to know how important it is to you to be able to stop the electronic dice yourself and not leave it to the computer to stop them. (It's just a matter of stopping the dice and not the possibility of selecting the probability of success or the project).

Thus, you will receive 30 new points. You can use a part or the totality of these points to purchase the right "to stop the electronic dice yourself". If you don't purchase this right, the dice are stopped in a random manner by the computer. If you do purchase this right, you will do the stopping yourself.

We are asking the following question:
Do you wish to pay to have the ability to stop the electronic dice yourself ?

Yes
No

If you click on "Yes", we will ask you to then indicate the maximum number of points you are ready to pay to have the ability to stop the electronic dice yourself (in the event that you have the right to decide).

When you are answering this question, have in mind the following process: you can purchase the right to stop the electronic dice yourself by indicating your maximum willingness to pay for this right – this must lie between 1 and 30. Then, a price between 1 and 30 will be drawn at random. If the price is less than or equal to your willingness to pay, you will pay the price and stop the electronic dice yourself. If the price is greater, you keep the 30 points and the electronic dice are stopped randomly. **With this procedure, it is best to honestly indicate how much you value the right to stop the electronic dice.**

Example 1: You are willing to pay a maximum of 5 points to have the ability to stop the electronic dice yourself (your readiness to pay is 5 points). The price that is randomly drawn is 18 points. As your readiness to pay is less than the price, you don't pay the price. You keep the whole 30 points and the electronic dice are stopped randomly.

Example 2: You are prepared to pay a maximum of 25 points to have the ability to stop the electronic dice yourself (your readiness to pay is 25 points). The price that is randomly drawn is 7 points. As your readiness to pay is greater than the price, you pay the price of 7 points. You keep 23 of the 30 points and stop the electronic dice yourself.

If you are ready to pay for the ability to stop the electronic dice, we ask that you indicate your exact readiness to pay.

At this moment in the experiment, you do not yet know which rounds will be drawn at random by the computer.
If you delegated the right to decide or did not have the right to decide in the round that was drawn randomly, you will not pay for the right to stop the electronic dice.

As well, if you chose the guaranteed payment in the two rounds drawn randomly from the second part of the experiment, you will not pay for the right to stop the electronic dice.

Please raise your hand if you have any questions about these instructions. We will come answer you at your seat. If not, click on the “CONTINUE” button.

D. Instructions for Loss Aversion Task (All subjects)

You now have the possibility to play in a series of lotteries. The potential earnings will be added to your total income, the potential losses will be subtracted from your total income.

You will soon be presented with a series of lottery decisions. For each lottery, please decide whether you accept or reject this lottery. At the end, one of the lotteries will be chosen at random.

If you accepted that lottery, a random process will determine whether you have won or lost the lottery.
If you rejected the lottery, nothing will happen and your total income will remain unchanged.

For each of the following lotteries, please choose whether to accept or reject the lottery:

1. With a probability of 50%, you win 5 euros; with a probability of 50%, you lose 1 euro.

2. With a probability of 50%, you win 5 euros; with a probability of 50%, you lose 2 euros.

3. With a probability of 50%, you win 5 euros; with a probability of 50%, you lose 3 euros.

4. With a probability of 50%, you win 5 euros; with a probability of 50%, you lose 4 euros.

5. With a probability of 50%, you win 5 euros; with a probability of 50%, you lose 5 euros.

6. With a probability of 50%, you win 5 euros; with a probability of 50%, you lose 6 euros.

E. Instructions for Cognitive Ability Task (All subjects)

(Raven's Progressive Matrices Test)

Information regarding the fourth part of the experiment

Please answer this last questionnaire, following the next instructions:

1. For each question, choose, among the 8 options shown on the bottom of the screen, the image most adapted to fill the black space in the picture above. In order to register your choice, click on the corresponding number in the right part of the screen, then click on the "OK" button.

2. There are 16 questions in total. Try to answer correctly to the most possible questions within a time limit of 10 minutes.

3. If you wish to reach one question directly, you can enter its number (1-16) and click on the "Go" button on the left part of the screen.

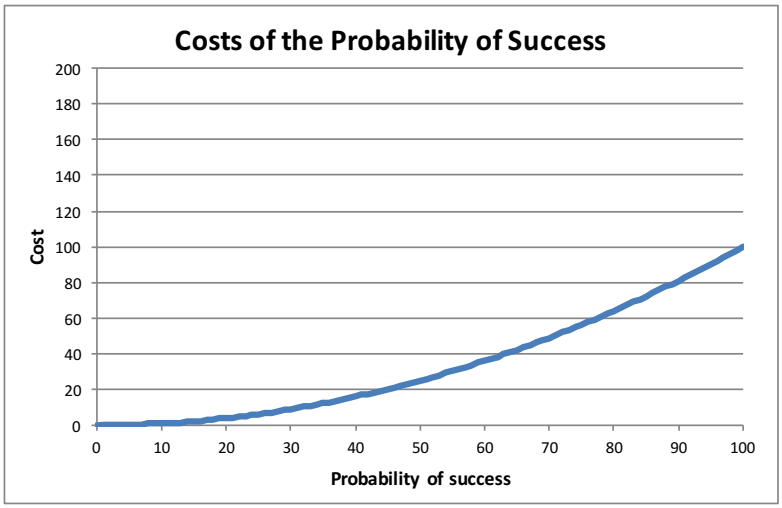
4. You can equally go to the preceding (next) question by clicking on the "Preceding" ("Next") buttons on the left low corner of the screen.

F. Supplementary Cost Sheet (All subjects)

(The figure and the table display all possible effort levels and their associated costs. This was distributed to all subjects to aid the determination of the intended effort level in Part 1 of the experiment. This is the "yellow" sheet with the cost parameter k=0.01. The "blue" sheet was equivalent with the respective costs with k=0.02)

Supplementary sheet with cost schedule

YELLOW SHEET



Probability of Success	Cost	Probability of Success	Cost	Probability of Success	Cost	Probability of Success	Cost
0	0						
1	0,1	26	6,8	51	26,1	76	57,8
2	0,2	27	7,3	52	27,1	77	59,3
3	0,3	28	7,9	53	28,1	78	60,9
4	0,4	29	8,5	54	29,2	79	62,5
5	0,5	30	9	55	30,3	80	64
6	0,6	31	9,7	56	31,4	81	65,7
7	0,7	32	10,3	57	32,5	82	67,3
8	0,8	33	10,9	58	33,7	83	68,9
9	0,9	34	11,6	59	34,9	84	70,6
10	1	35	12,3	60	36	85	72,3
11	1,3	36	13	61	37,3	86	74
12	1,5	37	13,7	62	38,5	87	75,7
13	1,7	38	14,5	63	39,7	88	77,5
14	2	39	15,3	64	41	89	79,3
15	2,3	40	16	65	42,3	90	81
16	2,6	41	16,9	66	43,6	91	82,9
17	2,9	42	17,7	67	44,9	92	84,7
18	3,3	43	18,5	68	46,3	93	86,5
19	3,7	44	19,4	69	47,7	94	88,4
20	4	45	20,3	70	49	95	90,3
21	4,5	46	21,2	71	50,5	96	92,2
22	4,9	47	22,1	72	51,9	97	94,1
23	5,3	48	23,1	73	53,3	98	96,1
24	5,8	49	24,1	74	54,8	99	98,1
25	6,3	50	25	75	56,3	100	100

General Conclusion

The main content of this dissertation is arranged in five chapters that intend to be self-contained pieces in themselves. For that reason, at the end of each chapter I have tried to discuss some of the insights that may emerge from each of the different analysis. Even so, a few tentative ending insights related to future research and the behavioral implications of multiple preferences can be spelled out.

First, Chapters 1 and 3 suggest that the time dimension of preferences is an important component for the study of social and economic behavior. They suggest that some of the problems that are traditionally treated as static, both in descriptive and welfare analysis, are best thought as dynamic. For example, Chapter 3 illustrates how time may be necessary for the rationalization of behavior that results from some decision making models based on endogenous and multiple preferences. Chapter 1 gives some examples of how excluding time from welfare analysis may create difficulties in ranking different states of affairs.

The time dimension is also essential for understanding the evolving and higher-order preferences discussed in Chapter 1. As a second insight, I am now convinced that more attention should be given to the endogenous evolution of preferences, taking into consideration how past preferences at $t-1$ (or earlier) relate to present preferences at t (or after). This may be mediated, among other relevant mechanisms, by the conflict between first- and second-order preferences. Among the several multiple preference models reviewed in the General Introduction of this dissertation, I now believe that the *evolving preferences model* with hierarchical preferences is an understudied and promising line of research.

This view of agency is an interesting perspective to look at questions related to personal identity, justice, well-being, and moral responsibility that I have only (if something) surfaced in this dissertation. Different views over personal identity can lead us to very different insights in terms of how to consider a state of affairs just, how to measure well-being, or how to impute moral responsibility to an action. I find it an exciting prospect for future research to develop some of the insights gathered in this dissertation into these broader (and in many ways more meaningful) topics of research.

Another insight, mostly connected to Chapters 1 and 2 of this dissertation, is the pertinence of using non-choice data in economic analysis. In Chapter 2 I have discussed some of the *potential* difficulties of relying exclusively on choice data. As I have argued there, relying exclusively on it will not always be problematic. But as I tried to motivate in both Chapters 1 and 2, using data such as stated choices, preferences, or other attitudes may be useful (and important) for both positive and normative analysis.

In fact, I find that survey-based controlled experiments are an interesting way to gather insights into some of the questions left open in this dissertation. It is important to be aware, and try to offset if possible, the potential weaknesses of such data. As argued in Chapters 1 and 2, the non-incentivized nature of most non-choice data creates potential difficulties in terms of attention and honesty. Ways to try to countervail these issues, include, for example, recording the duration taken for an answer in order to exclude speedy or random answers. Likewise, combining survey-based methods with choice-based decisions may be an interesting methodological extension for the topics of research I have mentioned so far.

Finally, the last two chapters suggest that both the introduction of multiple preferences to more applied settings (such as political behavior) and their empirical study is a challenging endeavor. I have left many open questions directly related to the content of these chapters. For example, future research could look if betrayal aversion, as defined in Chapter 4, is an empirical relevant phenomenon. Similarly, simpler

experimental designs may look at how the intrinsic value of decision rights relates with beliefs about the behavior of others, and if decision making models based on multiple preferences are good explanatory models of this kind of behavior.

I have wandered around some ways in which multiple preferences could be a relevant tool for social and economic analysis. Now is time to look back, and try going farther.

Caminante, son tus huellas
el camino y nada más;
Caminante, no hay camino,
se hace camino al andar.
Al andar se hace el camino,
y al volver la vista atrás
se ve la senda que nunca
se ha de volver a pisar.
Caminante, no hay camino
sino estelas en la mar.

Wayfarer, the only way
Is your footprints and no other.
Wayfarer, there is no way.
Make your way by going farther.
By going farther, make your way
Till looking back at where you've wandered,
You look back on that path you may
Not set foot on from now onward.
Wayfarer, there is no way;
Only wake-trails on the waters.

Antonio Machado, *Proverbios y Cantares*

Introduction (in French)

Introduction Général

L'objectif de cette introduction est d'abord de motiver la recherche menée dans cette thèse (section 1.1), en second lieu de conceptualiser la notion de préférences multiples (section 1.2), puis de donner un aperçu des chapitres de recherche (section 1.3), et, enfin, discuter brièvement des méthodes et la vision épistémologique adoptée dans cette recherche (section 1.4). Une partie de la littérature connexe est discutée dans le cadre de cette introduction, en particulier dans la section 1.2. La liste des références est fournie à la fin de l'introduction.

Motivation

Fernando Pessoa, l'un des écrivains portugais les plus prolifiques, et dans mon avis le plus brillant d'entre eux, a écrit sous le nom de plusieurs personnages de fiction qu'il avait créés. Ces personnages étaient, selon lui, plus que des pseudonymes. Ils étaient, au lieu de cela, ses « hétéronymes », dotés de leurs propres biographies, apparences, sentiments et visions du monde. Ils ont écrit mieux ou pire en portugais, sur des sujets différents, et dans différents styles.

Si la plupart d'entre nous ne reconnaissent pas ces personnages indépendants en nous-mêmes, l'idée que les personnes sont polyvalentes a une longue tradition dans la pensée philosophique. On peut retracer au moins à Platon l'idée que les êtres humains

sont divisés intérieurement. Selon Platon, l'affrontement entre le raisonnement moral et les passions immorales humaines était central :

> "First the charioteer of the human soul [reason] drives a pair, and secondly one of the horses is noble and of noble breed [moral], but the other quite the opposite in breed and character [passions]."
>
> Plato, Phaedrus

Les motivations conflictuelles ont continué à être un sujet dans la pensée philosophique pour les siècles à suivre. Par exemple, il existe une preuve écrite considérable selon laquelle, au cours des dix-septième et dix-huitième siècles, les auteurs se sont concentrés sur le conflit entre la morale, les passions humaines immorales et la poursuite de l'intérêt matériel qui, jusque-là, a été déprécié en tant que passion immorale de l'avarice (voir Hirschman 1977). Mais ce n'est qu'au XIXe siècle que les préférences multiples, sous la forme d'identités multiples, deviennent un sujet d'étude. William James a d'abord conceptualisé la notion de « plusieurs soi-même » comme suit :

> "Properly speaking, *a man has as many social selves as there are individuals who recognize him* and carry an image of him in their mind. To wound any one of these his images is to wound him. But as the individuals who carry the images fall naturally into classes, we may practically say that he has as many different social selves as there are distinct *groups* of persons about whose opinion he cares. He generally shows a different side of himself to each of these different groups. Many a youth who is demure enough before his parents and teachers, swears and swaggers like a pirate among his "tough" young friends. We do not show ourselves to our children as to our club-companions, to our customers as to the laborers we employ, to our own masters and employers as to our intimate friends. From this there results what practically is a division of the man into several selves ; and this may be a discordant splitting, as where one is afraid to let one set

> of his acquaintances know him as he is elsewhere ; or it may be a perfectly harmonious division of labor, as where one tender to his children is stern to the soldiers or prisoners under his command."
>
> William James, The Principles of Psychology

De nos jours, la psychologie et l'économie comportementale ont fourni des preuves empiriques considérables suggérant que le comportement est souvent dû ou s'explique par des motivations conflictuelles, des identités multiples ou les différents rôles que les gens mènent dans leur vie. Par exemple, certaines expériences suggèrent qu'être renvoyé à une des deux identités (l'identité asiatique ou américaine des sujets américano-asiatiques) déclenche différentes réponses comportementales en termes de patience (Benjamin et al. 2010) et de coopération (LeBoeuf et al. 2010). De même, certaines expériences suggèrent qu'être renvoyer à des constructions sociales liées à l'intelligence telles que « un professeur » ou « Albert Einstein » (par opposition à « un super model » ou « Claudia Schiffer ») affectent l'intelligence perçue de soi-même, son concept de soi, et le comportement subséquent en termes de résultats de tests (Dijksterhuis et al. 1998 ; Schubert and Hafner 2003 ; LeBouef and Estes 2004). En outre, l'accumulation de preuves ainsi que l'observation et l'introspection occasionnelles indiquent que le comportement de choix est souvent le résultat de préférences endogènes, c'est-à-dire des préférences qui dépendent de l'expérience du décideur.

Cependant, l'approche néoclassique dominante dans l'économie est de synthétiser les goûts, les valeurs, les intérêts et les objectifs individuels dans une relation de préférence unique, stable et exogène. Selon ce point de vue, l'identité personnelle d'un individu ne peut pas changer selon le contexte ou au fil du temps. Il n'y a pas de conflit interne qu'un individu n'est pas capable de résoudre, et aucune évolution ne sous-tend son expérience pendant le temps.

Le comportement de choix est, dans le modèle de choix néoclassique, supposé résulter de la maximisation de cette relation de préférence stable et exogène. Les implications observatoires de ce modèle sont décrites par les axiomes de préférence ré-

vélés, tels que les axiomes fort et faible des préférences révélées (voir Sen 1971 pour une revue de littérature). Dans certaines conditions, ces axiomes sont nécessaires et suffisants pour décrire un ensemble de choix comme si résultant de la maximisation d'une préférence stable et exogène.

Cette approche est problématique pour au moins deux raisons. Tout d'abord, il peut être difficile en termes de description et de prédiction du comportement économique. En particulier, les exigences de rationalité exigées par le modèle de choix néoclassique ne sont pas compatibles avec les comportements en raison de l'évolution des préférences, de plusieurs modèles d'apprentissage et d'autres déterminants contextuels et sociaux du comportement. Deuxièmement, cela pourrait égarer les économistes en termes d'inférence sur le bien-être et de classement du bien-être des différents états sociaux. D'une part, les préférences et les choix échouent souvent à révéler le bien-être des individus, car ils peuvent, entre autres, être le résultat d'une dissonance cognitive, d'une erreur flagrante ou d'une manipulation. Mais il est également possible, comme on le verra dans le chapitre 1, que le modèle de préférence unique évite les informations normatives pertinentes sur ce que les gens *apprécient*, ce dont ils se *soucient* et *qui* ils souhaitent être ou devenir.

Une autre vue du modèle de choix rationnel traditionnel est de supposer que l'agent économique est guidé par des préférences multiples. Selon cette vue, le comportement de choix n'est pas le résultat de la maximisation d'une préférence unique, mais plutôt du résultat de l'agrégation, du conflit ou du changement entre les relations de préférence multiples. Beaucoup de modèles de prise de décision qui intéressent les économistes et que la théorie standard ne peut expliquer, y compris les préférences changeantes et la formation des préférences, sont dus ou peuvent s'expliquer par des préférences multiples, identités ou plusieurs soi-même. De même, l'évolution des préférences des individus selon leurs expériences peut être due où s'expliquer par des préférences multiples au fil du temps.

L'objectif de cette thèse est d'explorer des modèles de prise de décision basés sur des préférences multiples comme alternative au paradigme de préférence unique.

Dans la première partie de la thèse, j'explore certaines des implications (comportementales) de l'adoption de préférences multiples en économie. Je révise certaines des conséquences positives et normatives de cette proposition (chapitre 1), la distinction comportementale entre les modèles de préférences uniques et multiples (chapitre 2) et présente un nouveau cadre de choix avec le temps dans lequel les modèles de préférences changeantes peuvent être plus facilement caractérisé (chapitre 3). La deuxième partie de la thèse est consacrée à l'analyse théorique et empirique du comportement économique qui peut être représenté comme s'il résulte de la prise de décision avec des préférences multiples. En particulier, je construis un modèle pour étudier les effets des préférences multiples sur le comportement politique (chapitre 4) et je mène une étude expérimentale pour distinguer les différentes motivations derrière une potentielle valeur intrinsèque du droit de décision (chapitre 5).

Avant de procéder à un aperçu des chapitres de recherche, je discute et conceptualise la notion de préférences multiples et fournit une taxonomie de modèles de préférences multiples qui peuvent être utiles pour contextualiser la recherche menée dans cette thèse et indiquer les futurs axes de recherche.

Préférences multiples

Il existe maintenant de nombreux modèles de prise de décision basés sur des préférences multiples qui sont utilisées pour expliquer le comportement économique. Un exemple est donné par les théories du double processus ou du système dual qui sont maintenant éminentes en économie (Thaler and Shefrin 1981, Bernheim and Rangel 2004, Fudenberg and Levine 2006, parmi beaucoup d'autres). L'hypothèse centrale partagée par tous ces modèles est que certains comportements économiques sont le résultat de l'interaction de deux types de prise de décision, une basée sur une délibération raisonnée / réflective et une autre sur les décisions impulsives/automatiques. Ces modèles sont utilisés pour expliquer des comportements économiques pertinents

tels que l'addiction (Bernheim and Rangel 2004) et le choix inter-temporel (Thaler and Shefrin 1981 ; Fudenberg and Levine 2006).

Préférences, Soi et Identité. Étant donné que les notions de préférences multiples, de multiples soi-même et d'identités multiples sont souvent utilisées et parfois interchangées dans la littérature, il est utile de préciser ce que je veux dire par chacune de ces notions. J'utilise le terme de *préférences multiples* comme un concept "parapluie", qui comprend une collection d'ordres (basée par exemple sur des motivations, des préoccupations ou des points de vue différents), des multiples soi-même ou des identités multiples. Je considère que les préférences sont ou peuvent être considérées comme une expression de soi, et que les différents modèles de prise de décision indiquent différentes notions sous-jacentes de l'identité personnelle de l'agent économique. J'utilise le terme *plusieurs soi-même* pour se référer aux cas où les préférences multiples sont modélisées comme des "sous-agents" qui interagissent les uns avec les autres comme s'ils étaient des joueurs dans un jeu interpersonnel. Enfin, j'utilise le terme *identités multiples* pour désigner les différentes identifications (sociales) que les individus peuvent retenir pour différents groupes ou adopter dans différents contextes.

Une taxonomie des modèles de préférences multiples. Les modèles de prise de décision basés sur des préférences multiples peuvent être distingués selon de nombreux critères. Sur la base de mes discussions précédentes et d'autres études (par ex. Ambrus and Rozen 2013), certains critères plausibles pour différencier les modèles de préférences multiples sont les suivants : (i) si toutes les préférences sont actives à chaque période, (ii) si ces préférences sont ou non stables au fil du temps, (iii) et si les préférences multiples sont indépendantes ou commensurables en une seule préférence à chaque période. Le premier critère distingue les modèles qui prennent le comportement à chaque période à la suite de la maximisation (ou autre processus) de l'une des préférences multiples, de ceux qui modèlent le comportement à la suite de l'agrégation (ou autre processus) des préférences multiples à chaque période. Le deuxième critère distingue les modèles qui prennent des préférences stables de ceux

qui n'assument aucune restriction *a priori* en termes de cohérence temporelle. Enfin, le troisième critère distingue les modèles qui assument un seul classement des alternatives à chaque période par rapport à ceux qui modélisent les différentes préférences comme des ordres indépendants, soi-même, ou identités.

Le tableau 1 présente une taxonomie provisoire de différents modèles de préférences multiples issus de l'intersection de ces trois critères. Dans ce qui suit, je discute chacune de ces représentations de l'agent économique avec une brève référence avec la littérature, à l'exception de la *préférence statique* qui n'est rien d'autre que le modèle de choix rationnel traditionnel.

Table 5.1 – Modèles de préférences multiples

	Stable	**Not Stable**
Préférence unique	Préférence statique	Préférences évolutives
Toutes préférences actives	Préférences simultanées	Préférences successives
Une (de beaucoup) active	Préférences alternées	

Préférences évolutives. Cette représentation conceptualise l'agent économique comme s'il était doté d'une identité personnelle qui évolue avec le temps. Cela représente l'individu en tant qu'agent en évolution qui prend ses décisions en fonction d'une séquence (endogène) de préférences multiples.

En économie, la plupart des modèles compatibles avec cette vue supposent une séquence exogène de préférences. Par exemple, le modèle de choix inter-temporel de Gul and Pesendorfer (2001, 2004, 2005) et le modèle de préférences changeantes qui est caractérisé au chapitre 3 sont compatibles avec l'évolution exogène des préférences. Il convient de noter que la psychologie, la philosophie et les neurosciences soutiennent la vision selon laquelle l'identité d'une personne évolue avec le temps.

Comme on l'affirme dans le chapitre 1, ce processus d'évolution peut être représenté par des préférences sur des préférences, également appelées préférences hiérarchiques, méta-préférences ou préférences de second ordre. En économie, certains auteurs ont préconisé l'utilisation de préférences sur des préférences (par ex. Sen 1977 ; Hirschman 1984), et dans le chapitre 1 j'esquisse deux modèles hiérarchiques

qui peuvent servir à des analyses économiques positives et normatives. La représentation (ou non) des préférences hiérarchiques est encore une autre distinction significative entre les différents modèles de préférences multiples.

Préférences simultanées. Cette représentation capte les décideurs qui ont une collection de préférences indépendantes, stables et actives à chaque période. Les modèles de ce type représentent l'agent économique comme s'il était doté d'une collection de préférences simultanées, c'est-à-dire d'une pluralité d'identités, de motivations, de points de vue ou de préoccupations distincts qui sont réparés au fil du temps.

Par exemple, le modèle de *pseudo-rationalisation* de Aizerman and Malishevski (1981), l'un des premiers à fournir des propriétés observables d'un modèle basé sur des préférences multiples, appartient à cette catégorie. Pour chaque situation de choix, l'agent sélectionne l'union des éléments maximaux de toutes les préférences, c'est-à-dire les éléments «meilleurs» pour au moins une préférence. Plus récemment, Cherepanov et al. (2013) s'appuient sur une représentation similaire pour proposer un modèle testable dans lequel un agent utilise une collection de préférences (interprété comme des histoires différentes qu'un agent se dit à lui-même) pour rationaliser un sous-ensemble d'options qu'il peut choisir. Toute option est rationalisable dans ce sens si elle est au moins meilleure pour l'une des préférences de l'agent (c.-à-d., Les options rationalisables sont l'union des éléments maximaux de toutes les préférences, celles sélectionnées dans Aizerman et Malishevski 1981). Ensuite, pour chaque situation de choix, l'agent choisit parmi ces options, celle qui est la plus préféré (maximal) selon une préférence unique, stable et exogène. Les modèles de choix par procédures séquentielles ou lexicographiques, tels que Tversky (1969), Manzini and Mariotti (2007, 2012), et Apesteguia and Ballester (2013) appartiennent également à cette catégorie. Dans ces modèles, dans toutes les situations de choix, un nombre arbitraire de préférences est appliqué séquentiellement pour sélectionner une alternative à choisir.

Préférences successives. Ce qui distingue les modèles basés sur des préférences successives de ceux basés sur des préférences simultanées est que pour la première, l'ensemble des préférences n'est pas nécessairement stable dans le temps. Dans les modèles de prise de décision basés sur des préférences successives, il existe un nouvel ensemble de préférences multiples actives à chaque période donné.

Par exemple, plusieurs des modèles « dual-self » du choix inter-temporel supposent qu'un « long-run self » interagit avec des successifs « short-term selves ». Fudenberg and Levine (2006, 2012), par exemple, modélisent un agent rationnel doté d'un soi-même stable et clairvoyant (un « planificateur ») qui interagit avec un nouveau soi-même myope (un « faiseur ») à chaque période (Fudenberg and Levine 2006) ou après un ensemble de périodes (Fudenberg and Levine 2012).

Préférences alternées. Cette représentation conceptualise l'agent économique comme s'il était dotée d'une collection de préférences (stables ou instables) et qu'il alterne entre elles d'une période à l'autre. La différence de cette représentation par rapport à celle des préférences simultanées est qu'une seule préférence, pas deux ou plus, dictera la décision à chaque période.

Par exemple, la théorie fondée sur la raison développée par Dietrich and List (2013, 2016) est conforme à cette représentation. Des modèles de préférence aléatoire, tels que Becker et al. (1963), Barberà and Pattanaik (1986), McFadden and Richter (1990), Loomes and Sugden (1995), Gul and Pesendorfer (2006), Apesteguia et al. (2017), entre autres, peuvent également être interprétés comme base sur les préférences alternées. Dans ces modèles, la préférence individuelle qui est active dans une situation de choix donnée est tirée au hasard d'un ensemble de préférences potentielles.

La vision selon laquelle les gens se comportent souvent d'une manière « single-minded », même si on les considère comme une collection de préférences, est partagée, par exemple, par Schelling (1984) et Gigerenzer and Selten (2000). Selon Schelling (1984), les gens sont mieux représentés comme un ensemble de « centres de valeurs » qui partagent les mêmes croyances et les mêmes capacités de raison-

nement, mais diffèrent en termes de volition. Selon ce point de vue, un centre de valeur (ou soi-même) agira comme un dictateur à chaque période, en remportant « le concours intime pour l'auto-commande » à cette période (voir Schelling 1984, 57-81). Selon Gigerenzer and Selten (2000), les indices dans l'environnement désigneront l'une des nombreuses heuristiques d'un agent. Étant donné que ces heuristiques avancent dans une fin particulière, les agents agissent selon un seul critère et d'une manière unique à chaque situation de choix, ce qu'on appelle « single-minded ».

Un aspect important qui distingue ces modèles est que les modèles de préférences évolutives traitent l'agent comme une unité d'agence (comme le modèle de choix rationnel traditionnel basé sur une préférence statique), alors que les modèles de préférences simultanées, successives et alternées représentent l'agent économique comme divisé entre plusieurs ordres, identités ou soi-même qui contestent le concours interne pour l'auto-commande. Cela distingue deux grandes représentations de l'agent économique en fonction des préférences multiples : (i) un **agent évolutif** qui décide selon l'évolution d'une préférence unique, et (ii) un **agent déchiré** qui décide en fonction du conflit entre plusieurs préférences. Bien qu'il existe maintenant une littérature économique approfondie sur les modèles basés sur des agents conflictuels qui désagrègent l'unité d'agence d'une personne en multiples ordres, identités ou soi-même, il a été fait moins d'efforts pour modéliser un agent évolutif qui prend des décisions en fonction d'une identité personnelle qui change selon son expérience dans le temps.

Vue d'ensemble

La recherche menée dans cette thèse est divisée en cinq chapitres. Je mets en place le cadre scientifique au chapitre 1, avec une évaluation de certaines des conséquences positives et normatives de l'adoption de modèles basés sur des préférences multiples en économie. Ce cadre se démarque du modèle de choix rationnel traditionnel et des modèles comportementaux récents qui traitent un comportement incompatible

avec la maximisation d'une préférence stable comme étant erronée. Je soutiens que, au lieu d'éviter un changement de préférence "authentique", il est important de distinguer les erreurs d'un comportement incohérent qui résulte des préférences (ou changement de préférence) auxquelles les individus s'identifient. Il s'agit de cas de préférences réflexives (auto-authentifiées) même s'ils contredisent la maximisation d'une préférence stable et exogène. J'introduis deux modèles hiérarchiques qui représentent certaines de ces idées et discute de la manière dont ils peuvent se rapprocher des modèles d'agent en conflit et d'agent évolutif afin de représenter un changement de préférence réflexif et non réflexif. Je soutiens que la distinction entre les préférences réflexives et non réflexives peut conduire à une meilleure description et à une meilleure prédiction du comportement économique, et je soutiens également que collecter des données sur lesquelles les préférences des individus s'identifient peut s'avérer utile pour l'économie normative, en particulier comme un perfectionnement des classements du bien-être actuellement utilisés dans l'économie comportementale du bien-être.

Dans les chapitres 2 et 3, je développe cette analyse au sein de la théorie du choix. L'interprétation commune donnée à un comportement de choix qui satisfait les axiomes de préférences révélées est qu'il résulte de la maximisation d'une relation de préférence stable et exogène. Au chapitre 2, je montre que l'observation des choix ne sont pas suffisantes pour exclure la possibilité qu'un comportement satisfaisant les axiomes de préférence révélées résulte plutôt de l'agrégation d'une collection de préférences distinctes. En particulier, je montre que tout ordre est équivalent d'une manière observationnelle à une agrégation majoritaire d'une collection d'ordres dichotomiques. Je montre également que tout ordre est équivalent d'une manière observationnelle à l'agrégation de Borda d'une collection d'ordres linéaires. J'utilise ces deux exemples et des résultats connexes pour discuter indissociabilité observationnelle et la sélection des modèles. Je soutiens que la question de l'indissociabilité peut s'étendre à des contextes où certains comportements de choix peuvent résulter d'une décision individuelle ou collective ; cependant je défends aussi que des ques-

tions concernant la plausibilité de différents modèles explicatifs et s'il est important d'identifier le modèle sous-jacent de décision doivent être posées avant de considérer l'indissociabilité théorique problématique. Dans le cas où l'indissociabilité est en effet problématique, une question reste ouverte sur les méthodes - en plus des données subjectives – qui devraient être utilisées pour identifier le modèle sous-jacent de la prise de décision.

Dans le chapitre 3, qui repose sur le travail en commun avec Nicolas Gravel, nous introduisons un cadre pour l'analyse des choix lorsque le dernier dépend explicitement du temps. Nous rapprochons ce cadre du cadre théorique traditionnel intemporel et illustrons son utilité en proposant trois modèles possibles de décision dans un tel cadre : (i) changement de préférences, (ii) formation de préférence par essai et erreur, et (iii) choix avec un biais endogène de *status-quo* en raison de l'inertie dans les préférences. Nous proposons une caractérisation complète de chacun de ces trois modèles de choix au moyen d'axiomes de préférence révélée qui ne peuvent être formulés dans un cadre intemporel. Bien que seul le premier d'entre eux soit rationalisé par un modèle de prise de décision basé sur des préférences multiples, notre analyse suggère le potentiel de ce cadre pour étudier d'autres modèles de choix motivés par des préférences endogènes et multiples.

Le chapitre 4 est consacré à la relation entre les identités multiples et le comportement politique, et repose sur un travail conjoint avec Sacha Bourgeois-Gironde. Nous développons un modèle spatial de comportements de participation et de vote qui cherche à expliquer le comportement des électeurs en conflit, c'est-à-dire des électeurs qui s'identifient à deux groupes ou parties. Ce sont les électeurs qui ont deux préférences conflictuelles (en tant qu'identités) et que, d'après l'aversion pour trahir l'une de leurs identifications, souhaitent satisfaire les deux préférences. Dans ces conditions, nous montrons que s'il n'y a pas de position qui réconcilie les vues idéologiques des deux parties, il est toujours rationnel que les électeurs en conflit s'abstiennent. Cela étant, même s'ils pouvaient, en tant que groupe, influencer le résultat de l'élection, nous l'appelons une malédiction de l'électeur en conflit. Dans

une compétition électorale à deux candidats, cette malédiction implique que les candidats convergent vers les résultats préférés des électeurs en conflit si et seulement si ces électeurs sont pivots et les parties partagent des points de vue idéologiques. Sinon, nous montrons que des équilibres convergents et divergents sont possibles en fonction du degré de polarisation des partis et si les candidats ont une idéologie ou non. Ces résultats illustrent comment le comportement de certains électeurs avec des préférences multiples ou identifications peut influencer les résultats électoraux et suggère que davantage de recherches devraient porter sur les électeurs mixtes et modérés qui composent le centre politique.

Enfin, le chapitre 5 est basé sur un travail en commun avec Nobuyuki Hanaki et Benoît Tarroux, et étudie les multiples motivations empiriquement. Nous concevons une étude expérimentale qui distingue les différentes motivations qui donnent lieu à une préférence pour le contrôle dans une interaction « principal-agent ». En particulier, nous perfectionnons une expérience récente de Bartling et al. (2014), dans laquelle ils ont constaté que les individus suisses attachent une valeur intrinsèque significative pour décider par eux-mêmes plutôt que de le déléguer à une autre personne. Nous introduisons une série de traitements afin de décomposer cette valeur entre (i) une préférence pour l'indépendance des autres, (ii) un désir de pouvoir, ou (iii) d'autres motifs tels qu'une préférence pour l'autosuffisance. En outre, nous effectuons une comparaison interculturelle entre la France et le Japon pour éclairer les déterminants sociaux de ces préférences. Nos principaux résultats suggèrent que (i) les individus japonais et français évaluent intrinsèquement les droits de décision au-delà de leurs avantages instrumentaux, que (ii) cette valeur est plus grande pour les Français que les Japonais, et que (iii) l'autosuffisance est la seule raison d'être de la valeur intrinsèque des droits de décision en France et au Japon. Nous avons également une légère preuve selon laquelle, bien que les principaux français soient indifférents en ce qui concerne l'indépendance et le pouvoir en tant que motifs de la valeur intrinsèque de leur contrôle, ils sont évalués négativement par les principaux japonais. Bien que notre expérience ne soit pas conçue de telle sorte que nous soyons

en mesure de déterminer si chaque individu est motivé par plus d'une préférence / motivation indépendante, cela suggère que ce pourrait être le cas pour les individus japonais qui semblent valoriser intrinsèquement l'autosuffisance et l'indépendance et pouvoir négativement.

Méthodologie de recherche

Dans cette partie de l'introduction, je présente brièvement la vision épistémologique adoptée dans cette thèse et les deux principales méthodes, analyses théoriques et expérimentales utilisées dans les cinq chapitres. Je commence par ces derniers.

Méthodes

La recherche menée dans cette thèse repose principalement sur deux méthodes : (i) la construction de modèles théoriques et (ii) la réalisation d'expériences de laboratoire.

Analyse théorique. Le raisonnement théorique et la modélisation ont une longue tradition en économie. Le raisonnement à travers les modèles présente plusieurs avantages. Par exemple, Walliser (2007) soutient qu'un modèle comporte six fonctions : l'emblématique (contextualisation, symbolisation et interprétation du phénomène économique dans un langage rigoureux), la syllogistique (explication, inférence et simulation du phénomène économique), l'empirique (la confrontation et la validation des idées théoriques contre les données empiriques), l'heuristique (stabilisation et évolution du savoir), la praxéologique (instrument de prédiction et ensemble d'action) et la rhétorique (expression concise, vulgarisation et transmission du savoir).

Dans le même temps, les modèles peuvent être fortement réduits et parfois peu liés à la réalité. Selon certaines écoles de pensée, il s'agit d'une déficience importante des modèles économiques. Avec l'avènement des grandes données et d'autres développements empiriques, le raisonnement théorique et la modélisation semblent être devenus moins importants au cours des dernières années. Comme défendu après, je

crois que les modèles théoriques sont utiles pour recueillir des visions du comportement économique.

Analyse expérimentale. La deuxième méthode utilisée dans cette thèse est la conception et la réalisation d'une analyse expérimentale de laboratoire. Cette méthode présente l'avantage (comparé à d'autres méthodes empiriques) de créer un environnement contrôlé qui est adapté pour isoler et étudier un ensemble limité d'effets et de relations causales. Une expérience en laboratoire « is a simple and controlled mini-world in contrast to the complex and uncontrolled maxi-world » (Maki 2005, 306). En conséquence, l'expérimentation peut apporter des informations précieuses sur la façon dont les gens se comportent, sur ce qu'ils apprécient, et comment et pourquoi ils le font.

Un inconvénient potentiel de la méthode expérimentale est la faible validité externe possible des résultats, c'est-à-dire leur faible applicabilité aux contextes du « monde réel » (voir par ex. Guala 1999 ; Loewenstein 1999 ; Starmer 1999). Par exemple, sans savoir pourquoi certains résultats comportementaux ont été obtenus dans un cadre expérimental, il peut être difficile d'utiliser les résultats au-delà du contexte où l'expérience a été exécutée. Dans le cas d'essais contrôlés randomisés (ECR), un type spécifique d'expériences de terrain couramment utilisées dans les paramètres de développement, Deaton (2010, 448) soutient que pour un ECR « to produce "useful knowledge" beyond its local context, it must illustrate some general tendency, some effect that is the result of mechanism that is likely to apply more broadly ». L'auteur soutient que les expériences devraient être axées sur la théorie et donne l'exemple de nombreuses expériences d'économie comportementale. Pourtant, l'étape du laboratoire à d'autres contextes peut être difficile pour d'autres raisons. Par exemple, il peut être difficile de créer des circonstances « parallèles » en laboratoire à la partie spécifique du système économique que l'expérience est destinée imiter.

Prenant en considération les avantages et les réserves de l'analyse expérimentale, les expériences en laboratoire semblent particulièrement adaptées aux cas où il est difficile d'isoler ou d'identifier les effets ou les relations occasionnelles d'intérêt dans

le monde réel. La séparation des valeurs instrumentales et intrinsèques associées à un comportement économique ou social, comme on l'a essayé au chapitre 5, semble correspondre à un tel cas. Les expériences de laboratoire semblent particulièrement appropriées, selon ce point de vue, pour éclairer les multiples motivations qui sous-tendent les comportements économiques et sociaux pertinents et qui sont difficiles à démêler dans d'autres contextes.

Déclaration épistémologique

Ces dernières années, un débat fructueux a entouré la question de l'épistémologie des modèles économiques (voir par ex. Maki 1994 ; Sugden 2000 ; Rubinstein 2006 ; Gilboa et al. 2014). Une dimension importante de cette discussion a été la définition de ce qui constitue un bon modèle. Par exemple, Rubinstein (2006, 881) soutient qu'un bon modèle théorique est comme une fable, c'est-à-dire un parallèle simplifié (éventuellement irréaliste) à une situation du monde réel qui « identifies a number of themes and elucidates them ». Selon Rubinstein (2006), les modèles ne sont pas censés être vérifiés et ont une portée limitée. Ils n'influencent pas le monde réel par des conseils judicieux ou une capacité prédictive, mais plutôt par une « accepted collection of ideas and conventions that influence the way people think and behave » (Rubinstein 2006, 882). Sugden (2000), d'autre part, considère qu'un bon modèle théorique est un monde crédible, c'est-à-dire une réalité parallèle au monde réel qui, compte tenu de notre connaissance des « general laws governing events in the real world », pourrait lui-même être accepté comme réel. Selon Sugden (2000), « the gap between model world and real world can be filled by inductive inference ». Selon ce point de vue, il est important de « recognize some significant similarity between those two worlds » (Sugden 2000, 23).

Modèles comme cannes à pêche. Mon point de vue se situe entre ces deux conceptions. La perspective prise dans cette thèse est qu'un modèle (ici défini comme un cadre théorique ou expérimental) est un outil pour former des *représentations*. Par représentation, je veux dire le résultat (i) d'une compréhension plus précise ou intui-

tive de la nature d'un effet ou d'une relation causale qui peut être observée dans le monde réel, ou (ii) une compréhension plus précise ou intuitive de certaines notions économiques ou du monde réel lui-même. Une telle représentation pourrait être une « tendance générale » dans le sens de la citation de Deaton (voir la section précédente), mais aussi une compréhension de certaines lois générales par la contextualisation, la symbolisation et l'interprétation du phénomène économique dans une configuration rigoureuse, soit théorique ou expérimental. Par exemple, les cadres expérimentaux sur les comportements électoraux et les identités sociales ont apporté des représentations (sous la forme d'une compréhension plus précise) de la nature des effets potentiels des identités sociales sur le comportement de vote dans le monde réel (par ex. Schram and Sonnemans 1996 ; Feddersen et al. 2009 ; Bassi et al. 2011).

Je crois que définir le *but* d'un modèle comme la collecte de représentations conceptuelles convient à l'analyse positive en économie. Selon cette perspective, un modèle (théorique ou expérimental) qui n'apporte aucune représentation est un modèle inutile. Par exemple, bien qu'il soit crédible, il semble en principe possible (bien que peu probable) qu'un modèle soit dépourvu d'idées pertinentes pour le monde réel. De même, un modèle qui a une finalité différente que de rassembler des représentations est, selon cette perspective, un modèle potentiellement utile mais peut-être inadéquat. Par exemple, une fable qui élucide un effet comportemental donné, mais qui a pour but principal de communiquer, déguiser, un point de vue moral semble être un modèle inadéquat.

De nombreux modèles sont également *contextuels*, dans le sens où les relations causales dans le monde réel sont plutôt relatives qu'absolues. Cela semble être le cas pour des modèles comportementaux positifs, dès que l'on prend une perspective mondiale ; plusieurs études expérimentales interculturelles ont documenté des différences significatives dans les préférences et le comportement au sein des différentes sociétés (voir le chapitre 5 pour les références). D'autres, comme les modèles d'identité personnelle, peuvent se rapporter à une sorte de représentation absolue. Mais

avoir à l'esprit le cadre contextuel des modèles semble être un outil utile d'interprétation.

Un bon modèle, selon cette perspective, est celui qui apporte des représentations pertinentes pour un groupe de personnes donné, un contexte et / ou un temps spécifique. Il peut s'agir soit d'une fable, soit d'un monde crédible. À mon avis, adopter une fable ou un monde crédible pour élucider un sujet donné, dépend de ce qui est le plus adapté comme outil pour recueillir des représentations sur ce sujet. Sugden (2000) fait valoir, à mon avis, que la crédibilité et l'inférence inductive peuvent favoriser cet objectif. Mais une fable (non crédible) peut être encore utile pour recueillir des représentations lorsque, par exemple, celles-ci permettent des compréhensions intuitives du monde réel lui-même.

Enfin, je crois que, tout comme les modèles « matériels » (expérimentaux) sont testés, certains prémisses et prédictions des modèles « théoriques » peuvent et doivent être testés. En particulier, il est possible d'apporter des preuves supplémentaires si ces connaissances sont valables pour un groupe de personnes, un contexte et/ou un temps spécifique. En effet, le processus de construction et de test des modèles semble être combiné et synergique.

Je crois que cette perspective modeste est adéquate en économie compte tenu de la tendance à juger les modèles (positifs) en fonction de leur cohérence empirique et de leur capacité prédictive, et de la fréquente surestimation de la capacité descriptive et prédictive de ces modèles. Dans cette perspective, l'économiste est le pêcheur, le modèle est sa canne à pêche, et les représentations sont les meilleures prises qu'il peut espérer pêcher.

Bibliographie

Aizerman, M. A. and Malishevski, A. V. (1981). General theory of best variants choice : Some aspects. *IEEE Transactions on Automatic Control*, 26(5) :1030–41.

Ambrus, A. and Rozen, K. (2013). Rationalising choice with multi-self models. *The Economic Journal*, 125 :1136–56.

Apesteguia, J. and Ballester, M. A. (2013). Choice by sequential procedures. *Games and Economic Behavior*, 77(1) :90–99.

Apesteguia, J., Ballester, M. A., and Lu, J. (2017). Single crossing random utility models. *Econometrica*, 85(2) :661–74.

Barberà, S. and Pattanaik, P. K. (1986). Falmagne and the rationalizability of stochastic choices in terms of random orderings. *Econometrica*, 54(3) :707–15.

Bartling, B., Fehr, E., and Herz, H. (2014). The intrinsic value of decision rights. *Econometrica*, 82(6) :2005–39.

Bassi, A., Morton, R. B., and Williams, K. C. (2011). The effects of identities, incentives, and information on voting. *The Journal of Politics*, 73(2) :558–71.

Becker, G., DeGroot, M., and Marschak, J. (1963). Stochastic models of choice behavior. *Behavioral Science*, 8 :41–55.

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

Bernheim, B. D. and Rangel, A. (2004). Addiction and cue-triggered decision processes. *The American Economic Review*, 94(5) :1558–90.

Cherepanov, V., Feddersen, T., and Sandroni, A. (2013). Rationalization. *Theoretical Economics*, 8(3) :775–800.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48 :424–55.

Dietrich, F. and List, C. (2013). Where do preferences come from ? *International Journal of Game Theory*, 42(3) :613–37.

Dietrich, F. and List, C. (2016). Reason-based choice and context-dependence : An explanatory framework. *Economics and Philosophy*, 32(2) :175–229.

Dijksterhuis, A., Spears, R., Postmes, T., Stapel, D., Koomen, W., van Knippenberg, A., and Scheepers, D. (1998). Seeing one thing and doing another : Contrast effects in automatic behavior. *Journal of Personality and Social Psychology*, 75(4) :862–71.

Feddersen, T., Gailmard, S., and Sandroni, A. (2009). Moral bias in large elections : Theory and experimental evidence. *American Political Science Review*, 103(2) :175–92.

Fudenberg, D. and Levine, D. K. (2006). A dual self model of impulse control. *The American Economic Review*, 96 :1449–76.

Fudenberg, D. and Levine, D. K. (2012). Timing and self-control. *Econometrica*, 80(1) :1–42.

Gigerenzer, G. and Selten, R. (2000). *Bounded Rationality : The Adaptive Toolbox*. MIT Press, Cambridge, MA.

Gilboa, I., Postlewaite, A., Samuelson, L., and Schmeidler, D. (2014). Economic models as analogies. *The Economic Journal*, 124 :513–33.

Guala, F. (1999). The problem of external validity (or "parallelism") in experimental economics. *Social Science Information*, 38 :555–73.

Gul, F. and Pesendorfer, W. (2001). Temptation and self-control. *Econometrica*, 69(6) :1403–36.

Gul, F. and Pesendorfer, W. (2004). Self-control and the theory of consumption. *Econometrica*, 72(1) :119–58.

Gul, F. and Pesendorfer, W. (2005). The revealed preference theory of changing tastes. *The Review of Economic Studies*, 72(2) :429–48.

Gul, F. and Pesendorfer, W. (2006). Random expected utility. *Econometrica*, 74(1) :121–46.

Hirschman, A. O. (1977). *The Passions and the Interests : Political Arguments for Capitalism Before its Triumph*. Princeton University Press, New Jersey.

Hirschman, A. O. (1984). Against parsimony : Three easy ways of complicating some categories of economic discourse. *The American Economic Review : Papers and Proceedings*, 74(2) :89–96.

LeBoeuf, R. A., Shafir, E., and Bayuk, J. B. (2010). The conflicting choices of alternating selves. *Organizational Behavior and Human Decision Processes*, 111(1) :48–61.

LeBouef, R. A. and Estes, Z. (2004). "fortunately, i'm no einstein" : Comparison relevance as a determinant of behavioral assimilation and contrast. *Social Cognition*, 22(6) :607–36.

Loewenstein, G. (1999). Experimental economics from the vantage point of behavioural economics. *Economic Journal*, 109(453) :25–34.

Loomes, G. and Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, 39 :641–48.

Maki, U. (1994). Isolation, idealization and thruth in economics. In *Idealization VI : Idealization in Economics, Poznan Studies in the Philosophy of the Sciences and the Humanities*, volume 38, pages 147–68. Rodopi, Amsterdam.

Maki, U. (2005). Models are experiments, experiments are models. *Journal of Economic Methodology*, 12(2) :303–15.

Manzini, P. and Mariotti, M. (2007). Sequentially rationalizable choice. *The American Economic Review*, 97(5) :1824–39.

Manzini, P. and Mariotti, M. (2012). Choice by lexicographic semiorders. *Theoretical Economics*, 7 :1–23.

McFadden, D. and Richter, M. K. (1990). Stochastic rationality and revealed stochastic preference. In Chipman, J. S., McFadden, D., and Richter, M. K., editors, *Preferences, Uncertainty, and Optimality : Essays in Honor of Leo Hurwicz*, pages 163–186. Westview Press : Boulder, CO, 161-186., Boulder, Colorado.

Rubinstein, A. (2006). Dilemmas of an economic theorist. *Econometrica*, 74(4) :865–83.

Schelling, T. (1984). *Choice and Consequence : Perspectives of an Errant Economist*. Harvard University Press, Cambridge, MA.

Schram, A. and Sonnemans, J. (1996). Why people vote : Experimental evidence. *Journal of Economic Psychology*, 17 :417–42.

Schubert, T. W. and Hafner, M. (2003). Contrast from social stereotypes in automatic behavior. *Journal of Experimental Social Psychology*, 39(6) :577–84.

Sen, A. K. (1971). Choice functions and revealed preferences. *Review of Economic Studies*, 38 :307–17.

Sen, A. K. (1977). Rational fools : A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs*, 6(4) :317–44.

Starmer, C. (1999). Experiments in economics : Should we trust the dismal scientists in white coate ? *Journal of Economic Methodology*, 6 :1–30.

Sugden, R. (2000). Credible worlds : The status of theoretical models in economics. *Journal of Economic Methodology*, 7(1) :1–31.

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

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76(1) :31–48.

Walliser, B. (2007). Les fonctions des modèles économiques. In Leroux, A. and Livet, P., editors, *Leçons de Philosophie Économique Tome III : Science Économique et Philosophie des Sciences*, pages 285–302. Economica, Paris.

Abstract

In this thesis I explore decision making models based on multiple preferences. In the first part of the thesis, I analyze some of the implications of adopting multiple preferences in economics and different ways in which they can be conceptualized and used within this field. In particular, I review some of the positive and normative consequences of preferences over preferences (Chapter 1), the behavioral (in)distinguishability of the single and multiple preferences models (Chapter 2), and introduce a new framework of choice with time in which models of changing preferences can be more easily characterized (Chapter 3). The second part of the thesis is devoted to the theoretical and empirical analysis of economic meaningful behavior that can be represented as if it is the result of decision making with multiple preferences. In particular, I build a model to study the effects of multiple preferences to political behavior (Chapter 4), and run an experimental study to distinguish different motivations behind a potential intrinsic value of holding a decision right (Chapter 5).

Keywords: Multiple preferences; Behavioral welfare economics; Revealed preference theory; Reflexive preferences; Preference change; Time; Spatial voting; Conflicted voters; Intrinsic value; Decision rights; Cross-cultural experiment.
JEL classification: B4; C91; D01; D03; D6; D7; P16.

Résumé

Dans cette thèse, j'explore les modèles de prise de décision basés sur des préférences multiples. Dans la première partie de la thèse, j'analyse certaines des implications de l'adoption de préférences multiples en économie et de différentes façons dont elles peuvent être conceptualisées et utilisées dans ce domaine. En particulier, je révise certaines des conséquences positives et normatives des préférences sur des préférences (chapitre 1), la distinction comportementale entre des modèles de préférences uniques et des modèles de préférences multiples (chapitre 2), et j'introduis un nouveau cadre de choix avec le temps dans lequel les modèles de préférences multiples peuvent être plus facilement caractérisés (chapitre 3). La deuxième partie de la thèse est consacrée à l'analyse théorique et empirique du comportement économique qui peut être représenté comme s'il résulte de la prise de décision avec des préférences multiples. En particulier, je construis un modèle pour étudier les effets des préférences multiples sur le comportement politique (chapitre 4) et je mène une étude expérimentale pour distinguer les différentes motivations derrière une potentielle valeur intrinsèque du droit de décision (chapitre 5).

Mots-Clés: Préférences multiples; Théorie des préférences révélées; Préférences réflexives; Changement de préférences; Économie comportementale du bien-être; Temps; Vote spatial; Électeurs en conflit; Valeur intrinsèque; Droits de décision; Expérience interculturelle.
Classification JEL: B4; C91; D01; D03; D6; D7; P16.

Croyez ceux qui cherchent la vérité, doutez de ceux qui la trouvent.

André Gide, *Ainsi soit-il ou Les jeux sont faits*