THÈSE

Université de Toulouse

En vue de l’obtention du

DOCTORAT DE L’UNIVERSITÉ DE TOULOUSE

Délivré par : l’Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue le 02/09/2013 par :

Frédéric MOISAN

The Bonds of Society: an Interdisciplinary Study of Social Rationality

JURY

PIERPAOLO BATTIGALLI  Professeur - Université de Bocconi  Rapporteur
JAN VAN EIJCK  Professeur - ILLC, Université d’Amsterdam  Rapporteur
GIUSEPPE ATTANASI  Maître de Conférences - Université de Strasbourg  Examinateur
JAN-GEORG SMAUS  Professeur - Université Paul Sabatier  Président
ASTRID HOPFENSITZ  Maître de Conférences - TSE  Co-directrice
EMILIANO LORINI  Chargé de Recherche - IRIT, CNRS  Co-directeur
INGELA ALGER  Chargé de recherche - TSE, CNRS  Invitée
FRANÇOIS SCHWARZENTRUBER  Maître de Conférences - ENS-Cachan, IRISA  Invité

École doctorale et spécialité :

MITT : Domaine STIC : Intelligence Artificielle

Unité de Recherche :

Institut de Recherche en Informatique de Toulouse

Directeur(s) de Thèse :

Andreas HERZIG, Astrid HOPFENSITZ et Emiliano LORINI

Rapporteurs :

Pierpaolo BATTIGALLI et Jan VAN EIJCK
A ma Mamie
A few years ago, it is my curiosity about science that naturally led me to engage in this very exciting adventure. Since my engineering days, I have always had a keen interest in using mathematics and logic to better understand human thinking and behavior. Today, I can say that studying in Toulouse was the best choice I could have made to pursue my doctoral study. In fact, during the last four years, I have had the chance and privilege to work with great people. Writing this dissertation would have never been possible without their invaluable help, their precious ideas, and their enthusiasm for discussions.

Primarily, I would like to express my deepest gratitude to my wonderful supervisors, Andreas Herzig, Astrid Hopfensitz, and Emiliano Lorini, without whom this thesis would clearly not exist. I am particularly grateful to Andreas and Astrid for bringing together their complementary skills and ideas in computer science and economics to initiate this interdisciplinary project. I also thank them for trusting me to carry out such a challenging project. Doing multidisciplinary research has been for me an extremely rewarding experience and I now believe more than ever that the fields of computer science and economics have a lot to gain from sharing their methodologies.

Moreover, I would like to thank Andreas for playing a major role in convincing me that logic is a very adequate tool to formalize and understand human behavior. I have always appreciated his interest about everything and his availability for discussions about any topic. I have particularly enjoyed our discussions about the logic of social emotions. Similarly, I want to thank Astrid for transmitting me her keen interest in behavioral and experimental economics. She has taught me everything that I know about designing and running economic experiments. I am particularly grateful that she involved me in her TIES project, which has largely contributed to the development of this thesis. She has indeed convinced me that social relationships are not only important to our lives, they are also very interesting to study. More generally, it is always a real pleasure for me to come by her office and discuss about economic and psychological matters among other things.
I would like to offer my special thanks to Emiliano who quickly became my main thesis advisor. In fact, he is the only person who has been involved in all my research work so far. I can fairly say that our collaboration has come naturally given the extent of our common interests in logic, game theory, and behavioral sciences. By encouraging me to look into various disciplines such as computer science, economics, psychology, and philosophy, he has taught me that research has no boundaries. More generally, it has been for me a lot of fun working with him on all those exciting topics, and I really thank him for that. I have also highly appreciated his passion for research, his patience (good research requires time!), as well as his honest comments and advices regarding my work.

In addition, I am grateful to all my PhD advisors for providing me with the opportunity to travel to various conferences in order to present my work. Through those events, I have had the chance to visit some of the most beautiful places in the world, including Ljubljana, Malaga, Sevilla, Osuna, Stockholm, and New York (twice).

Among other people involved in this thesis, I would like to thank Giuseppe Attanasi for his important collaboration regarding our experimental study. I especially appreciated his precious help in running what was then my first experiment.

I am also particularly grateful to the volleyball coach at the Toulouse 1 University Capitole, Patrick Galard, who allowed us to run our experiment with his players during a training session. Clearly, a large part of this thesis would not exist without his precious collaboration.

Furthermore, I am indebted to Pierpaolo Battigalli and Jan Van Eijck for examining my thesis and providing both kind comments and crucial remarks about it. Similarly, I thank Ingela Alger, François Schwarzen-truber, and Jan-Georg Smaus for their highly relevant comments as well as their very entertaining questions during the defense of this thesis.

I also want to thank Martine Labruyère for nicely taking care of the administrative procedure that led to the defense of this thesis. Things have become much easier through her precious help.

Over the last four years, I have not only done research, and so I want to express my deepest gratitude to Sylvie Doutre and Laurent Perrussel who have been involved in supervising my teaching activities during all that time. It has always been a great pleasure working with them,
and more generally, they really made me enjoy teaching, which, I have to say, was not an easy task. Similarly, I want to thank Jean-Marc Thévenin for continuously sharing his experience as well as his cheerfulness.

I also want to thank the students with whom I have had the chance to share my office at IRIT: Antoun Yaacoub and Naji Obeid who introduced me to the appealing Lebanese culture, and Farid Aiche with whom I particularly enjoyed discussing differences and similarities between Algerian and French societies (including their national football teams).

More generally, I would like to thank all the other great people I have had the pleasure to meet at IRIT and TSE during all those years in Toulouse. I have clearly learned a lot from them too.

Apart from my work environment, many people have contributed to make my PhD study such a nice experience.

In particular, I would like to thank every member of my cycling team US Colomiers Cyclisme for all the nice moments I have spent road racing with them during the last years in the Toulouse area. They have clearly strengthened my feeling that cycling is the most beautiful sport in the world.

I am grateful to Mr Milou who was an amazing bicycle mechanic I have had the privilege to meet at rue du midi in Toulouse. I think of him as the happiest man I have ever seen, and I will never forget everything he did for me.

I am also indebted to Saeko with whom I will always share very strong ties. I have sometimes wondered what I would be doing now if I had not met her Down Under years ago, and all I can say is that I would have then missed a lot of great opportunities, including writing this thesis. I cannot thank her enough for the unconditional and invaluable support she has offered me during all this time.

My final thanks go to my family, and particularly my parents who have always stood by my side during all these years of study despite their only partial understanding of what I was doing. I am thankful to my sister who has always provided me with her precious support whenever I needed it. I also want to thank them, along with my great aunt Yvette and my great uncle Jean, for coming all the way from Brittany to attend my PhD thesis defense in Toulouse. It is their presence that really made this day unforgettable to me.
Finally, I would like to dedicate this thesis to my grandmother, in memory of all the wonderful moments I have had the chance to spend in this little piece of paradise that we used to call “Chez Mamie”. I have always admired her extreme kindness, her exemplary simplicity, her extraordinary humility, her remarkable courage, and her very special attention towards others. She has always encouraged me to do whatever I wanted to do in life, and she has largely inspired me to study human behavior throughout this thesis. I will never forget our endless but so enjoyable conversations. For all that, and for so much more: Merci Mamie.

Toulouse, September 2013
Contents

1 Introduction 1

2 Rationality in Social Interactions 9
  2.1 Rational preferences 10
    2.1.1 Ordinal utility 10
    2.1.2 Cardinal utility 11
  2.2 Game theory 14
    2.2.1 Simultaneous games 14
    2.2.2 Sequential games 17
  2.3 Interactive epistemology 23
    2.3.1 Aumann’s common knowledge 24
    2.3.2 Epistemic modal logic 26
  2.4 Towards social rationality 36
    2.4.1 Experimental game theory 36
    2.4.2 Other-regarding preferences 42

3 Epistemic Rationality in Dynamic Games 47
  3.1 A modal logic of actions, strategies, knowledge and preferences 48
    3.1.1 Syntax 48
    3.1.2 Semantics 50
    3.1.3 Some validities 54
  3.2 Backward induction and rationality 54
    3.2.1 Nash equilibrium 54
    3.2.2 Backward induction 57
    3.2.3 Epistemic rationality 58
  3.3 A syntactic proof of Aumann’s theorem 59
  3.4 A more convenient characterization of knowledge 60
  3.5 Beyond Aumann’s theorem 61
  3.6 Related work in logic 69
6 Towards Collaborative Societies

6.1 Problems of co-operation .............................................. 147
6.2 Team reasoning ......................................................... 151
   6.2.1 Definitions ....................................................... 151
   6.2.2 Limitations ....................................................... 155
6.3 A theory of social ties .................................................... 156
   6.3.1 Definitions ....................................................... 157
   6.3.2 Analysis .......................................................... 160
   6.3.3 Relationship with team reasoning .............................. 160
      6.3.3.1 Gradual group identification ............................ 160
      6.3.3.2 Similarities in any two-player game .................... 162
      6.3.3.3 Further comparison in any $n$-player game ........... 164
6.3.4 Illustration: the Three Musketeers game ....................... 166
6.4 A theory of empathetic preferences ................................... 170
6.5 Conclusion ............................................................ 174

7 Summary and Future Research ........................................... 175

8 Résumé en français ......................................................... 182

8.1 Introduction .......................................................... 182
8.2 Rationalités dans les interactions sociales ....................... 186
8.3 La rationalité épistémique dans les jeux extensifs ............... 187
8.4 Les liens sociaux et la coordination stratégique ................ 188
8.5 Les effets comportementaux des liens sociaux ................... 192
8.6 Vers des sociétés collaboratives .................................... 193
8.7 Conclusion ............................................................ 195

A Proofs in modal logic ELEG ............................................. 198

A.1 Proof of Lemmas ....................................................... 198
   A.1.1 Syntactic proof of lemma A.1 ................................. 199
   A.1.2 Syntactic proof of lemma A.2 ................................ 201
   A.1.3 Syntactic proof of lemma A.3 ................................ 203
   A.1.4 Syntactic proof of lemma A.4 ................................ 204
   A.1.5 Syntactic proof of lemma A.5 ................................ 204
   A.1.6 Syntactic proof of lemma A.6 ................................ 205
   A.1.7 Syntactic proof of lemma A.7 ................................ 206
A.2 Proofs of Theorems .................................................... 209
   A.2.1 Syntactic proof of theorem 3.1 ............................... 209
   A.2.2 Syntactic proof of theorem 3.2 ............................... 210
   A.2.3 Syntactic proof of theorem 3.3 ............................... 212
   A.2.4 Syntactic proof of theorem 3.4 ............................... 212
B Instructions of the Experimental Study

B.1 Preliminary Instructions .......................................... 217
B.2 Instructions of the Baseline game ................................... 217
  B.2.1 Questions .......................................................... 219
B.3 Instructions of the Entrance game ................................. 219
  B.3.1 Questions .......................................................... 221
B.4 Belief questions ....................................................... 221
  B.4.1 In the Baseline game ............................................. 221
  B.4.2 In the Entrance game ........................................... 222
B.5 Elicited beliefs ....................................................... 223

C Social ties and team reasoning ..................................... 226

C.1 Proof of Theorem 6.2 .................................................. 226
C.2 Gradual group identification ........................................ 227
C.3 Proof of Theorem 6.3 .................................................. 229
C.4 Utility transformation vs. agency transformation ............... 231
C.5 Forming sub-coalitions in the Three Musketeers game .......... 232
  C.5.1 Prediction in a social ties game ............................... 233
  C.5.2 Prediction in an unreliable team interaction structure ....... 233

References .................................................................. 240
Chapter 1

Introduction

"Science does not know its debt to imagination."
— Ralph Waldo Emerson
Letters and Social Aims (1876)

“It is the spur of ignorance, the consciousness of not understanding, and the curiosity about that which lies beyond that are essential to our progress.”
— John Pierce

What does it mean to be rational? It is fair to say that this simple question has raised many debates over the past years and across various disciplines including philosophy, economics, psychology, and sociology. Yet, despite the difficulty for scientists to provide a clear definition, the term “rational” surprisingly appears to be overly used in everyday conversations, thereby indicating that people at least have a general idea of its concept. Indeed, everybody would agree that, broadly speaking, rationality involves thinking and behaving reasonably or logically. However, such a simplistic definition remains largely ambiguous as it simply pushes the problem away: what does it mean to think and behave reasonably? As an attempt to give a more precise answer to this question and investigate the role that such a principle plays in human cooperation, this dissertation combines various approaches and methodologies from computer science (logic) and economics (game theory, experimentations). However, in order to justify the need for such an interdisciplinary study, let us start by distinguishing between the concepts of reason and rationality: while one may indeed define reason through the psychological capacity for establishing and verifying facts based on perceived information, rationality is instead rather involved with the process of optimizing choices. In this case, as such conscious choices clearly result from some internal thinking, one can state that rationality naturally implies the use of reason.
Furthermore, it is customary among philosophers to distinguish between the notions of theoretical and practical rationality (see e.g., Kalberg [1980]). On one hand, theoretical rationality, which relies on evidential and argumentative support, simply deals with regulating one’s own beliefs\(^1\). On the other hand, practical rationality corresponds to the strategy of living one’s best possible life, achieving one’s most important goals, and maximizing one’s own preferences in as far as possible.

In order to illustrate these philosophical concepts, suppose that I know that abundant smoking kills, and that it implies that if I smoke, it will certainly deteriorate my physical health. Assuming that I prefer to stay alive and in good health for as long as possible, then I would be practically irrational to smoke even one cigarette because this choice alone would then not lead me to be in the best possible health (i.e., it is not optimal). Alternatively, being practically rational to smoke would require me to reconsider either my beliefs (e.g., I do not believe that smoking will deteriorate my health) or my preferences (e.g., I somehow do not care to have a long-lasting life). On the other hand, I may be theoretically rational to believe the fact that smoking a single cigarette will only have a negligible effect on my health. However, note that it would be theoretically irrational to believe that it will not deteriorate my health at all (this would conflict with my initial belief of the opposite statement).

Following this intuitive example, it should be clear that human beings are not rational by definition, but they can think and behave rationally or not, depending on whether they apply the strategy of theoretical and practical rationality to the thoughts they accept and to the actions they perform. Moreover, it is worth noting that both of these concepts are mutually supportive. In fact, while theoretical rationality can clearly help me accomplish my practical aims, practical rationality can allow me to improve the quality of my beliefs. Given the high relevance of both of these notions of rationality to the functioning of human behavior, especially in the context of social interactions, they will therefore together characterize the main focus of this dissertation.

Over the last decades, investigating the role that practical rationality actually plays in the social world has become the primary goal of many economists whose aim has been to use mathematical tools to model decision making. This interest has naturally led to the development of the area of game theory, which represents the study of strategic decision making through mathematical models of conflict and cooperation between intelligent rational agents (Gintis [2000]; Myerson [1997]; Osborne and Rubinstein [1994]; Osborne [2004]). The concept of a game

\(^1\)To be more rigorous, acceptances should be distinguished from beliefs, as argued in Tuomela [2000]: although both concepts are cognitive states, beliefs are involuntary (i.e., they are not subject to direct voluntary control), whereas acceptances are voluntary and intentional.
such a theory refers to can theoretically represent any sort of social interaction. Formally, a (non-cooperative) game can be described as a set of players, a set of actions (or strategies) available to those players, and a specification of payoffs for each combination of actions. One should note that, although game theory was primarily rooted in economics, it is now also extensively used and studied in other fields such as political science, psychology, as well as in philosophy and biology. In particular, the interest has more recently expanded to the area of computer science through the growing development of artificial intelligence and multi-agent systems. Real world applications are manifold and include robotics, electronic communication networks (e.g., electronic commerce), interactive education and entertainment (i.e., human-computer interactions), and resolving problems in security and safety. In any such situations, in order for artificial agents to efficiently interact with human beings, they must clearly be able to understand the basic principles of social behavior, as it is elicited in human societies. Such a problematic therefore strongly suggests the need for a formal theory of rationality in the context of social interactions.

However, despite its undeniable relevance to the study of practical rationality, classical game theory does not appear to be the most efficient tool to investigate the underlying connections with theoretical rationality. In fact, as suggested before, theoretical rationality basically deals with following some consistent and optimal way of reasoning, and game theory does not provide a sufficiently rich language that allows to unambiguously model this sort of thinking. In order to meet the needs to formalize some logical thinking, the use of propositional logic is often considered by computer scientists. Such a formal system is indeed concerned with reasoning about propositions, each of which basically represents a possible state of the world: for example, given two propositions $p$ and $q$, if $p$ is true and it is the case that “if $p$ is true, then $q$ is true”, then it can be inferred that $q$ is also true (this inference rule is known as Modus Ponens). However, in return for being very simple, such a logic is also not sufficient to express relevant statements defining the various types of mental states that may be elicited by human beings in social interactions. For this purpose, an alternative formal system was introduced to extend this logic with the addition of extra operators expressing modalities of the sort “it is possible/necessary that . . .”, “it is permitted/obligatory that . . .”, “one believes/knows that. . .”, etc. . . These additional modal operators are indeed of particular interest to the study of rationality because they allow to formally express and reason about some agents’ mental attitudes. For example, one can define the following rule in such a formal language: if individual $i$ believes that

---

1 A game is non-cooperative in the sense that it represents a detailed model of all the moves available to the players, in contrast with cooperative games, which abstract away from this level of detail and describes only the outcomes that result when the players come together in different combinations.
proposition $p$ is true, then $i$ believes that $i$ believes that $p$ is true (this rule is usually known as positive introspection). This type of logic, which is known as modal logic (Blackburn et al. [2002]; Chellas [1980]; Hintikka [1962]; Hughes and Cresswell [1968]), is also often associated with other logics such as epistemic logic (reasoning about knowledge), temporal logic (reasoning about time), deontic logic (reasoning about obligations), and dynamic logic (reasoning about complex programs). More generally, applications of modal logic are particularly important in philosophy, linguistic, and various areas of computer science such as artificial intelligence, distributed systems, database theory, program verification, and cryptographic theory. One aim of this thesis is to show that it is also particularly relevant to the field of economics.

Nevertheless, although the combination of game theory and logic clearly represents a powerful analytical tool to theoretically investigate the essential principles of rationality, their only limitation lies in the highly idealized views they often offer. As suggested earlier, the main motivation for formalizing rationality is to be able to accurately predict human behavior in social interactions. In fact, being rational becomes useless if I mistakenly believe that other individuals are rational, which may eventually lead me to perform poor actions with possibly catastrophic consequences. It is now widely accepted that classical game theoretic models have indeed failed to their original motivation to accurately predict human decision-making, which explains the recent growing interest for economic experimentation. Similarly to the physical sciences, the use of controlled experiments has indeed proven to be highly relevant to study economic questions (Camerer [2003]; Roth and Kagel [1995]). In order to test the validity of some economic theories, such experiments usually use cash to motivate human subjects in order to mimic real-world incentives. Such empirical methods allow to explore very important concepts such as altruism, fairness, reciprocity, and emotions, which have all, for a long time, been ignored by classical economics theories. In fact, some extensive empirical evidence already suggests that human beings are genuinely other-regarding in social interactions. However, while such studies have allowed to clarify the way in which every individual can actually contribute to the promotion of a society (e.g., through fair and cooperative behavior), one may wonder about the impact that a given society can have on its own members. Indeed, any human society simply consists of a group of individuals related to each other through more or less persistent social relationships. It therefore seems reasonable to claim that one’s social environment is largely responsible for determining one’s own well-being. After all, one’s contribution to the welfare of a society has for main (if not only) purpose to improve one’s quality of life within that society.

The main contribution of this dissertation is therefore to investigate this issue by studying some crucial aspects of human rationality as it is actually exhibited
in the context of social interactions. We attempt to provide an explanation of how these factors can eventually lead to some social behavior that promotes the welfare of the society as a whole. For the purpose of obtaining a realistic formal definition of the complex concept of social rationality, this study relies on the various methodologies that were previously introduced: game theory, modal logic, and economic experimentation. More specifically, we will argue that the following components are essential to the definition of social rationality in interactive situations:

- **Individual preferences.** Since practical rationality deals with optimizing goals, one indeed needs to be able to express what is the best outcome for every given individual. The approach that we follow here is quantitative regarding this matter, that is, it relies on agents measuring their utility for every event that may occur.

- **Knowledge and beliefs.** It is clear that theoretical rationality largely depends on the epistemic state an agent may hold. Considering the above example, if I do not know that smoking kills, then it can be rational for me to smoke. In the context of social interactions, the problem becomes even more complex since all agents must also consider what they know about what each other knows.

- **Social bonds between individuals.** This factor is generally not mentioned in the literature when defining rationality. We here claim that the type of social relationships that may exist between individuals involved in some social interactions has some effect on their respective rationality. More precisely, we argue that the level with which each agent is tied with every other agent directly affects his individual preferences, and consequently his behavior.

More precisely, this dissertation is organized as follows.

In Chapter 2, we provide a brief overview of the existing formal theories underlying the concept of rationality. More precisely, we present utility theory along with classical game theory in order to represent rational preferences as well as rational decision making in the context of social strategic interactions. As a means to reason about knowledge and beliefs, we similarly introduce a description of modal epistemic logic along with a simple example that illustrates its expressive power. Furthermore, we demonstrate the need for extending classical game theory so that it can incorporate the types of reasoning that are carried out by human subjects in various economic experiments. Our main claim will be that the apparent evidence for irrational behavior may, in fact, often not conflict with the classical assumption of rationality.

Through Chapter 3, we investigate the epistemic foundations of rationality through a logical analysis of social sequential interactions. Although similar work
have already been done in economics (e.g., Aumann [1995, 1999a]; Aumann and Brandenburger [1995]), the particularity of this analysis is that it relies on formal modal logic. In fact, we argue that modal logic is an invaluable tool that allows in-depth analyses while expressing concepts that are either informally or vaguely claimed to be captured by the classical game-theoretic language. In particular, we show that such a formal tool is ideal to unambiguously model the agents’ knowledge about what each other knows. This study therefore refers to the quite recent subarea called “Formal Interactive Epistemology” (a term coined by Aumann in Aumann [1999b]), which deals with the logic of knowledge and belief when there is more than one agent. Furthermore, the other main characteristics of this work is that it considers the temporal dimension of rationality, which is often ignored in the literature. We therefore show that the large expressive power of the proposed logic allows to bring some insight about the existing relationship between time and knowledge in social interactions. In order to illustrate the use of such a logic, we provide a syntactic proof of Aumann’s well known theorem stating that backward induction in perfect information games can be derived from the assumption of common knowledge of mutual rationality (Aumann [1995]). As a result, we show that such a logical study not only allows to clearly identify the required epistemic assumptions that are only implicit in Aumann’s original proof of the theorem, but also leads to weaken its original statement and answer relevant questions related to the mechanisms of learning and recalling, positive and negative introspections, temporal reasoning and bounded rationality. Moreover, we show that such an analysis further allows to give a formal answer to the main criticism of Aumann’s theorem from the game theory literature (see Stalnaker [1998]).

Following this epistemic analysis of individual rationality, the next chapters then restrict their focus on assuming that human beings are not solely driven by pure individualism. In fact, over the last few decades, the failure of classical economic theory has naturally led to the development of the field of behavioral economics, which basically consists in providing alternative explanations to any observed deviation from an optimal individualistic behavior. Most existing work in this area supports the influence of relevant genuine factors such as fairness, altruism, reciprocity, trust, and emotions on rational behavior (e.g., Berg et al. [1995]; Güth et al. [1982]). However, while those theories clearly allow for a more realistic view of social interactions, it also assumes the following strong assumption: one’s behavior is independent of the identity of the persons that are involved in a given interaction. Indeed, is one’s behavior likely to be the same when one interacts with a perfect stranger as when one interacts with one’s best friend? The most intuitive answer to this question clearly appears to be negative.

It is therefore our aim through Chapter 4 and Chapter 5 to study the nature of social relationships (between, e.g., friends, married couples, family relatives,
colleagues, class mates, teammates, etc...) as a possible explanation of human cooperative behavior. More specifically, Chapter 4 consists of a theoretical analysis of social ties through the design of a particular type of two-player game, which, we show, allows to disentangle predictions from theories based on self-interest, social preferences (i.e., inequity aversion, fairness), and social ties. Such a study therefore suggests the need to introduce a novel theoretical model built upon the main hypothesis that such social relationships influence a player’s choice by modifying his preferences. In order to verify our theoretical predictions, we then present, in Chapter 5, the design of an experiment that is based on the previous game. Such an experimental study involves subjects who share some genuine bonds with one another (selected participants were members of a sport club). In addition to varying the strength of social ties by allowing multiple interactions with individuals from different groups, we further measure the influence of two different types of social ties: a subjective tie determining how one feels about a social tie, and an objective tie specifying what is the actual value of a social tie.

Finally, we present, in Chapter 8, a generalization of the previous model of social ties as a means to formally represent rational cooperative behavior in the context of strategic interactions possibly involving more than two individuals. We illustrate the advantage of this model by performing a detailed comparative analysis with another relevant theory from the economics literature that can similarly explain cooperation when agents act as members of the same group: Bacharach’s theory of team reasoning (Bacharach [1999]). As a result of this study, we demonstrate that the proposed model of social ties provides a simpler and more intuitive approach to modeling collaborative actions in the context of complex social interactions where competing groups may coexist.

One should note that the work presented throughout this dissertation are based on various important collaborations. More precisely, Chapter 3 is an extension of a joint work with Emiliano Lorini that has been published in Lorini and Moisan [2011]. This work was presented at the 4th workshop on Logical Aspects of Multi-Agent Systems (November 2011, Osuna, Spain). Chapter 4 is based on an article co-authored with Giuseppe Attanasi, Astrid Hopfensitz and Emiliano Lorini, which has been accepted for publication in the journal Phenomenology and the Cognitive Sciences. This work was also presented at the 2012 Social Networks and Multiagent Systems symposium (July 2012, University of Birmingham, UK), and at the Collective Intentionality VIII conference (August 2012, University of Manchester, UK). Furthermore, Chapter 5 corresponds to a joint work in progress with Giuseppe Attanasi, Astrid Hopfensitz and Emiliano Lorini, which was presented at the 2012 international ESA (Economic Science Association) conference.
(June 2012, New York University, USA) as well as at the international conference of game theory (July 2013, Stony Brook University, USA). Similarly, Chapter 8 represents a joint work in progress with Emiliano Lorini, which was presented at the seventh workshop in Decision, Game, and Logic (June 2013, KTH Stockholm, Sweden) and at the international conference of game theory (July 2013, Stony Brook University, USA).

It is also worth mentioning that, for the consistency of this dissertation, some other relevant published work have voluntarily not been included here: in an article co-authored with Andreas Herzig, Emiliano Lorini and Nicolas Troquard (Herzig et al. [2011]), we have presented a dynamic logic of propositional assignments that allows to represent and reason about normative systems. Moreover, in another joint work with Andreas Herzig and Emiliano Lorini (Herzig et al. [2012]), a similar simple logic has also been proposed to model the concept of trust (as it is theorized in Castelfranchi and Falcone [1998, 2010]; Falcone and Castelfranchi [2001]).
Chapter 2

Rationality in Social Interactions

“Do I really look like a guy with a plan? You know what I am? I’m a dog chasing cars. I wouldn’t know what to do with one if I caught it. You know, I just... do things.”

— The Joker (Heath Ledger)
The Dark Knight (2008)

“Scratch an altruist and watch a hypocrite bleed.”

— Michael Ghiselin

Through his famous claim that “man is a rational animal”, Aristotle asserts that rationality is an essential property of humankind, i.e., what distinguishes man from beast. While this principle has since been subject to many criticisms among philosophers, it has also been considered as a primary assumption in economics. More generally, everybody (including Aristotle himself) would agree that having the capacity for acting rationally does not prevent one from behaving irrationally. Instead, what the above principle suggests is that human beings have a natural tendency to use their reason as a means to seek and attain their highest possible level of welfare, thereby leading them to a perpetual “pursuit of happiness”.

Following this interpretation, we therefore present, through this chapter, an overview of the most relevant formal theories underlying the concept of rationality in the context of social interactions: a theory of measuring and representing rational preferences, a theory of strategic interactions, a theory of reasoning about knowledge and beliefs, and a theory of other-regarding interests.
2.1 Rational preferences

Preferences are obviously crucial in defining rationality as they basically represent what drives an individual towards a particular goal (i.e., what is most preferred). Moreover, such preferences play a particularly important role in this dissertation as we will later argue that they are partly shaped by some social factors. We therefore present, through this section, the two main economic approaches that allow to measure and formalize an individual’s preferences over a set of alternatives: a theory of ordinal utility, and a theory of cardinal utility.

2.1.1 Ordinal utility

Given an exhaustive set of mutually exclusive outcomes or consequences \( O = \{o_1, \ldots, o_n\} \) describing some possible states of the world (e.g., \( o_1 \) may state that “it rains today in Toulouse”), let us consider a ranking over \( O \) that defines one’s preferences. Formally, we introduce a weak preference relation \( \preceq \), which is defined as follows:

\[
\preceq: O \times O
\]

\( o_i \preceq o_j \) therefore reads “outcome \( o_j \) is at least as good as outcome \( o_i \).”

Alternatively, note that an individual’s preferences can also take other forms:

- **Strict preference of outcome** \( o_j \) **over outcome** \( o_i \) (noted \( o_i < o_j \)) occurs whenever \( o_j \) is weakly preferred to \( o_i \), and \( o_i \) is not weakly preferred to \( o_j \). Formally, this means that \( o_i < o_j \) if and only if \( o_i \preceq o_j \) and not \( o_j \preceq o_i \).

- **Indifference between two outcomes** \( o_i \) and \( o_j \) (noted \( o_i \approx o_j \)) occurs whenever \( o_j \) is weakly preferred to \( o_i \), and \( o_i \) is weakly preferred to \( o_j \). Formally, this means that \( o_i \approx o_j \) if and only if \( o_i \preceq o_j \) and \( o_j \preceq o_i \).

Moreover, in order to be rational, the preferences must at least satisfy the two following axioms:

- **Completeness**: all outcomes can be ranked in terms of preference. Formally, given two outcomes \( o_i \) and \( o_j \), we have that either \( o_i \preceq o_j \) or \( o_j \preceq o_i \) (or both in the case of indifference between \( o_i \) and \( o_j \)).

- **Transitivity**: all outcomes can be compared with other outcomes. Formally, given three outcomes \( o_i \), \( o_j \), and \( o_k \), if \( o_i \preceq o_j \) and \( o_j \preceq o_k \), then we have that \( o_i \preceq o_k \).

However, an individual’s preferences are often described by a utility function or payoff function. Thus, such a function, noted \( U \), specifies a real number assigned by the individual for each available outcome:
One can then state that such a utility function $U$ represents a preference relation $\preceq$ that satisfies the above axioms if and only if, for every two outcomes $o_i, o_j \in O$, $U(o_i) \leq U(o_j)$ implies $o_i \preceq o_j$. In this case, if $U$ represents $\preceq$, then it implies that $\preceq$ is complete and transitive.

As an illustration of this ordinal utility function $U$, suppose that I prefer outcome $o_1$ the most, followed by $o_2$, $o_3$, and $o_4$, that is $o_4 \preceq o_3 \preceq o_2 \preceq o_1$. If it is required to assign real numbers to these outcomes to reflect this ordering, then there are innumerable ways. One possible immediate assignment would be:

$$U(o_1) = 4; U(o_2) = 3; U(o_3) = 2; U(o_4) = 1$$

Clearly, there exists an uncountably infinite number of utility functions $U$ to define my preferences. Note that, while such an ordinal utility function captures the ranking of preferences, it says nothing regarding how much an outcome is more preferred over another one (i.e., the absolute or relative magnitude of preferences is ignored here). While determining such ordinal preferences can be achieved easily, it however does not characterize a very precise measure. We therefore consider an alternative approach that allows to fill this gap.

**2.1.2 Cardinal utility**

In contrast with an ordinal scale for measuring preferences, as presented in the previous section, let us now assume that not only the order of utilities is important, but the ratios of differences between utilities are also meaningful. Such an interpretation then refers to $U$ as a cardinal utility function. However, the problem of comparing the sizes of utilities does not appear to be an easy task. In fact, while I may say that eating an apple pie is preferable to eating a cheese cake, it is more difficult for me to say that it is twenty times preferable to the cheese cake. The reason is that the utility of twenty cheese cakes is hardly twenty times the utility of a cheese cake, by the law of diminishing returns.

This is why Von Neumann and Morgenstern suggested an extremely elegant theory for determining cardinal utilities in Von Neumann and Morgenstern [1944]. Their idea is indeed to consider lotteries instead of simple outcomes. More specifically, given a set of possible outcomes $O$ (as in the previous section), one can define a lottery $L$ as a discrete distribution of probability on $O$ (i.e., for every outcome $o \in O$, $L(o)$ specifies the probability that $o$ occurs). In this case, note that $\sum_{o \in O} L(o) = 1$.

For example, given $O = \{o_1, o_2\}$, the lottery $L$ defined by $L(o_1) = 0.75$ and $L(o_2) = 0.25$ denotes the scenario where the probability for outcome $o_1$ to occur
is 0.75 and that of outcome $o_2$ is 0.25.

Based on this concept, Von Neumann and Morgenstern introduce a way to rationally compare such lotteries. Indeed, they generalize the relation $\preceq$ from the previous section so that it now defines a preference ordering over lotteries. In this case, writing $L \preceq L'$ means that "$L'$ is at least as good as $L$" (i.e., "$L'$ is weakly preferred to $L$). Note that strict preferences and indifference between lotteries can also be expressed as before. In order for such preferences to be rational, they then propose the following set of axioms:

- **Completeness**: all lotteries can be ranked in terms of preferences. Formally, we have that, given two lotteries $L$ and $L'$ defined on the same set of outcomes $O$, either $L \preceq L'$ or $L' \preceq L$ (or both).

- **Transitivity**: preferences over lotteries are consistent. Formally, given three lotteries $L$, $L'$, and $L''$ defined on the same set of outcomes $O$, if $L \preceq L'$ and $L' \preceq L''$, then $L \preceq L''$.

- **Continuity**: there exists a "point" between being better than and worse than a given middle option. Given three lotteries $L$, $L'$, and $L''$ defined on the same set of outcomes $O$, if $L \preceq L'$ and $L' \preceq L''$, then there exists a probability $p \in [0, 1]$ such that $p \cdot L + (1 - p) \cdot L'' = L'$.

- **Independence**: it assumes that a preference holds independently of the possibility of an alternative outcome. Given two lotteries $L$ and $L'$ defined on the same set of outcomes $O$, for any lottery $L''$ (also defined on $O$) and any probability $p \in [0, 1]$, $L \preceq L'$ if and only if $p \cdot L + (1 - p) \cdot L'' \preceq p \cdot L' + (1 - p) \cdot L''$.

The Von Neumann-Morgenstern utility theorem is then defined as follows:

**Theorem 2.1** Given a set of outcomes $O$ and a preference relation $\preceq$ over lotteries defined on $O$ that satisfies completeness, transitivity, continuity, and independence, there exists a utility function $U : O \to \mathbb{R}$ such that, for any two lotteries $L$ and $L'$ defined on $O$:

$$L \preceq L' \text{ if and only if } EU(L) \leq EU(L')$$

where, for any lottery $L$, $EU(L)$ denotes the expected utility of $L$:

$$EU(L) = \sum_{o \in O} L(o) \cdot U(o)$$

Following this new concept of rational preferences, as it is usually called in economics, one can then specify an intuitive method for measuring cardinal utilities. As a means to illustrate such a procedure, let us now consider a simple
set of three outcomes $O = \{o_1, o_2, o_3\}$ such that the following preference ordering holds: $o_3 \prec o_2 \prec o_1$. The first step is to assign arbitrary utility values to both the most preferred outcome $o_1$ and the least preferred outcome $o_3$, with the only constraint that the utility of $o_1$ must be larger than that of $o_3$. For example, let us take $U(o_1) = 200$ and $U(o_3) = 100$. The difficult part comes with determining the utility of outcome $o_2$, which can be solved using Von Neumann and Morgenstern’s theory for rationally comparing lotteries, as shown in Theorem 2.1. In fact, let us consider the lottery $L$ that $o_2$ occurs with probability 1 (i.e., $L(o_2) = 1$ and $L(o_1) = L(o_3) = 0$). In this case, note that the expected utility of lottery $L$ simply corresponds to the utility of outcome $o_2$, that is, $EU(L) = U(o_2)$. Let us further define some lottery $L_1$ specifying that outcome $o_1$ occurs with probability $p$ (i.e., $L(o_1) = p$) and $o_3$ occurs with probability $1 - p$ (i.e., $L(o_1) = p$, $L(o_3) = 1 - p$, and $L(o_2) = 0$). One then only has to determine which lottery is preferred:

- if $L \prec L_1$, then it implies that the utility of outcome $o_2$ is strictly lower than the expected utility of lottery $L_1$, that is, $U(o_2) < p \cdot U(o_1) + (1 - p) \cdot U(o_3)$;
- if $L_1 \prec L$, then it implies that the utility of outcome $o_2$ is strictly higher than the expected utility of lottery $L_1$, that is, $U(o_2) > p \cdot U(o_1) + (1 - p) \cdot U(o_3)$;
- if $L \approx L_1$, then it implies that the utility of outcome $o_2$ is equal to the expected utility of lottery $L_1$, that is, $U(o_2) = p \cdot U(o_1) + (1 - p) \cdot U(o_3)$.

It is then straightforward to show that, by repeating this procedure through various meaningful values of probability $p$ (i.e., asking the same question as above while replacing $L_1$ with other lotteries), one can converge towards a realistic utility value for outcome $o_2$.

However, one might wonder whether such a cardinal utility function is a realistic representation of preferences. Are people really capable of expressing their preferences over lotteries? As any answer to this question would clearly be subject to debate, let us instead discuss on the actual need for such a function in the first place. In fact, it appears that cardinal utility has already proven its relevance in solving problems related to decision making under risk, representing inter-temporal preferences, or determining collective welfare evaluations. Throughout this thesis, we therefore follow this theory as a means to later express group preferences that are based on the members’ individual preferences. In particular, although it has been argued that aggregating cardinal utilities across persons is unrealistic (mainly because different people may indeed have different “zeros” on their utility scale), we will show in later chapters that such interpersonal comparison of cardinal utility is required when dealing with social concepts such as fairness, inequity aversion, and empathy.
2.2 Game theory

In this section, we present the most well known theory used to study and model rational decision making in social interactions. Game theory is a central topic in this dissertation because it represents a great tool to formally characterize the underlying principles of conflict and cooperation between intelligent agents. In order to provide a brief overview of the main concepts of game theory, we distinguish between two different types of games through the following sections: simultaneous and sequential games (see Gintis [2000]; Myerson [1997]; Osborne and Rubinstein [1994]; Osborne [2004] for more detailed introductions).

2.2.1 Simultaneous games

Let us consider the following situations. Two individuals, say Alice and Bob, have been arrested by the police as they are suspected for committing a crime together. However, the police admits there is not enough evidence to convict the pair on the principal charge. As a means to obtain some confession, the individuals are isolated from each other (i.e., they have no way to communicate with each other). The police then visit each of them individually and offer them the same following deal: the one who offers evidence against the other can be freed immediately. More precisely, if none of them agrees to betray the other, they are in fact cooperating with each other and against the police, and as a consequence, both of them will get to spend two years in jail, which corresponds to a rather limited punishment because of lack of proof. However, if only one of them betrays the other, by confessing to the police, the corresponding defector will be freed. The one who remains silent, on the other hand, will receive a severe punishment by having to serve ten years in prison, as a result of being uniquely held responsible for the crime. Of course, there is a catch to the story: if both prisoners testify against each other, both will be sentenced to five years in jail.

This particular scenario is called a simultaneous game because it involves individuals that have to make simultaneous decisions (i.e., both players have no information about the other’s choice). The question that naturally follows from this game is then: what should each individual tell the police? As a means to answer such a question, game theory actually offers an elegant way of representing the corresponding problem. In fact, let us consider the corresponding graphical representation of the above scenario depicted in Figure 2.1, where, for each individual $i$ (with $i$ referring to either Alice or Bob), actions $C_i$ and $D_i$ respectively stand for $i$’s choice to stay silent (i.e., cooperating with the other individual), and to confess to the police (i.e, betray the other individual).

Figure 2.1 therefore specifies both individuals’ utilities for every possible outcome, each of which consists of a combination of both agents’ actions. For exam-
ple, \((-10, 0)\) represents the outcome for Alice performing \(C_a\) and Bob choosing \(D_b\), which can be interpreted as follows: in the case of occurrence of this outcome, Alice’s utility is \(-10\) (i.e., she gets ten years in jail) while Bob’s is 0 (i.e., he gets to be freed).

Based on such a representation of the above situation, it then becomes straightforward to observe that the best action for both agents is to select \(D_a\) and \(D_b\). In fact, no matter what the other does, one appears to be always better off confessing to the police. This interpretation then leads to the following dilemma: while both agents would get five years in jail if they both optimally defect (i.e., by selecting \(D_i\)), it happens that it would be better for each of them to simultaneously choose \(C_i\) as they would then each get only two years in jail. We however postpone the more detailed analysis of such a game to Chapter 8.

Let us now present the general representation of any such simultaneous game, as shown through the following definition.

**Definition 2.1 (Standard Strategic Game)** A strategic game is a tuple \(G = \langle \text{Agt}, \{S_i| i \in \text{Agt}\}, \{U_i| i \in \text{Agt}\} \rangle\) where:

- \(\text{Agt}\) is a finite set of agents involved in the game;
- for every agent \(i \in \text{Agt}\), \(S_i\) defines the set of pure strategies available to \(i\);
- for every agent \(i \in \text{Agt}\), \(U_i : \prod_{j \in \text{Agt}} S_j \rightarrow \mathbb{R}\) defines a total payoff function mapping every combination of strategies to some real number.

Note, from Definition 2.1, that the concept of a pure strategy can be reduced to a simple action. A more complex definition of a pure strategy will however be introduced in the next section. As an illustration, the above situation can be represented as the strategic game \(G = \langle \text{Agt}, \{S_i| i \in \text{Agt}\}, \{U_i| i \in \text{Agt}\} \rangle\) where:

- \(\text{Agt} = \{Alice, Bob\}\);
- \(S_{Alice} = \{C_a, D_a\}\), \(S_{Bob} = \{C_b, D_b\}\);

<table>
<thead>
<tr>
<th></th>
<th>(C_b)</th>
<th>(D_b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_a)</td>
<td>(-2, -2)</td>
<td>(-10, 0)</td>
</tr>
<tr>
<td>(D_a)</td>
<td>(0, -10)</td>
<td>(-5, -5)</td>
</tr>
</tbody>
</table>

Figure 2.1: Prisoner’s dilemma


- $U_{\text{Alice}}(C_a, C_b) = U_{\text{Bob}}(C_a, C_b) = -2$, $U_{\text{Alice}}(D_a, D_b) = U_{\text{Bob}}(D_a, D_b) = -5$, $U_{\text{Alice}}(C_a, D_b) = U_{\text{Bob}}(D_a, C_b) = -10$, and $U_{\text{Alice}}(D_a, C_b) = U_{\text{Bob}}(C_a, D_b) = 0$.

Let us further abbreviate $S_J$ as the combination of strategies among agents in $J$, that is, $S_J = \prod_{i \in J} S_i$. In this case, we write $S$ instead of $S_{\text{Agt}}$, and $S_{-i}$ instead of $S_{\text{Agt}\setminus\{i\}}$.

In such strategic games, a solution concept then represents a formal rule that predicts how the game will be played. The most well known solution concept in the economics literature is the Nash equilibrium, which specifies a combination of strategies such that, no player has anything to gain by changing his own strategy unilaterally. The formal form of this equilibrium concept is found in Definition 2.2.

**Definition 2.2 (Pure strategy Nash equilibrium)** A pure strategy Nash equilibrium in a strategic game $G = (\text{Agt}, \{S_i|i \in \text{Agt}\}, \{U_i|i \in \text{Agt}\})$ is a combination of pure strategies $s \in S$ such that, for every $i \in \text{Agt}$:

$$U_i(s'_i, s_{-i}) \leq U_i(s_i, s_{-i}) \text{ for every } s'_i \in S_i$$

Applying this principle to the above prisoner’s dilemma from Figure 2.1, the only Nash equilibrium is for both players to defect (i.e., Alice chooses $D_a$ while Bob selects $D_b$). It is usually assumed in game theory, that such a Nash equilibrium reflects some optimal behavior that each individual can rationally aim at.

Moreover, one may extend this concept so that it also considers mixed strategies. A mixed strategy for some agent $i$ then simply corresponds to a randomization of $i$’s pure strategies from $S_i$. For example, in the above game, instead of choosing between the choices $C_a$ and $D_a$, Alice may throw a dice in order to determine her actual choice (e.g., choosing $C_a$ with probability $2/3$ and choosing $D_a$ with probability $1/3$). Formally, this leads us to define the mixed extension of a strategic game, as in Definition 2.3.

**Definition 2.3 (Mixed extension)** Given a strategic game $G = (\text{Agt}, \{S_i|i \in \text{Agt}\}, \{U_i|i \in \text{Agt}\})$, the mixed extension of $G$ is the strategic game $MG = (\text{Agt}, \{\Delta(S_i)|i \in \text{Agt}\}, \{EU_i|i \in \text{Agt}\})$ such that:

- for every agent $i \in \text{Agt}$, $\Delta(S_i)$ is the set of probability distributions over $S_i$;
- for every agent $i \in \text{Agt}$, an expected utility function $EU_i : \prod_{j \in \text{Agt}} \Delta(S_j) \to \mathbb{R}$ assigns to each combination of mixed strategies $\alpha \in \prod_{j \in \text{Agt}} \Delta(S_j)$, the expected value under $U_i$ of the lottery over $S$ that is induced by $\alpha$:

$$EU_i(\alpha) = \sum_{s \in S} \left( \prod_{j \in \text{Agt}} \alpha_j(s_j) \right) \cdot U_i(s)$$

16
Following this consideration of mixed strategies, one can therefore also define the concept of a mixed strategy Nash equilibrium.

**Definition 2.4 (Mixed strategy Nash equilibrium)** A mixed strategy Nash equilibrium in a strategic game $G = \langle \text{Agt}, \{S_i| i \in \text{Agt}\}, \{U_i| i \in \text{Agt}\}\rangle$ is a pure strategy Nash equilibrium in its mixed extension $MG$ according to Definition 2.3.

### 2.2.2 Sequential games

In the previous section, we considered simultaneous games that assume that the players do not observe each other’s actions. However, it is clear that such an assumption is unrealistic in many situations that involve individuals interacting sequentially with one another. As an illustration, let us consider the following scenario, which is often referred to as the ultimatum game in the economics literature. Suppose that two players, say Alice and Bob, have to decide how to divide a sum of money (e.g., 10€) that is given to them. Alice first proposes how to divide the sum between the two players, and Bob can then either accept or reject this proposal. For any proposal from Alice, if Bob rejects it, neither player receives anything. On the other hand, if Bob accepts it, the money is split according to the proposal.

Such a game is played sequentially because, when taking his decision, Bob may take into account Alice’s previous choice. In other words, Bob has a full observation over Alice’s decision at the time of his decision.

As shown in Figure 2.2, such an interactive situation can be graphically represented by a game tree, where each vertex defines a decision node, each edge defines a particular action, and each leaf corresponds to an outcome of the game.

Figure 2.2 represents a binary version of the above ultimatum game where the players only have binary options. In this case, vertex $v_0$ defines Alice’s decision node while vertices $v_1$ and $v_2$ define Bob’s. Vertices $v_3$, $v_4$, $v_5$, and $v_6$ denotes the terminal nodes of the game where no player can make any decision (i.e., they express the outcomes of the game). Furthermore, actions $C_a$ and $D_a$ respectively stand for “Alice suggests to split equally the amount” and “Alice suggests to split the amount in her favour”. Similarly, $C_b$ and $D_b$ respectively stand for “Bob accepts Alice’s proposal” and “Bob rejects Alice’s proposal”.

Looking at the game representation from Figure 2.2, one can observe that, at both vertices $v_1$ and $v_2$, Bob is always better off playing $C_b$. In response to this behavior, Alice should then select $D_a$ at $v_0$. In other words, this prediction reflects the only solution that survives common knowledge of both players’ rationality in the ultimatum game.

As a means to perform a more detailed analysis of such sequential games, let us now present its general formal representation, as shown through the following
Figure 2.2: Binary ultimatum game

Definition 2.5 (Extensive game with Perfect Information) An extensive game with perfect information is a tuple $EG = \langle Agt, V, EndV, Q, A, next, \{U_i| i \in Agt\} \rangle$ where:

- $Agt$ is a finite set of agents involved in the game;
- $V$ is a finite set of vertices defining the nodes of the game tree;
- $EndV \subseteq V$ is the set of terminal nodes;
- for every vertex $v \in V \setminus EndV$, $Q(v) \in Agt$ defines the agent that moves at $v$;
- for every vertex $v \in V \setminus EndV$, $A(v)$ defines the set of actions available at $v$;
- for every vertex $v \in V \setminus EndV$ and every action $a \in A(v)$, $next(v, a) \in V$ defines the successor node that can be reached by performing $a$ at $v$;
- there exists a unique initial vertex $v_0 \in V \setminus EndV$ such that, for every $w \in V \setminus EndV$ and every $a \in A(w)$, $next(w, a) \neq v_0$;

$^1$We here restrict our analysis to extensive games with perfect information as they are the primary concern of the following chapters of this dissertation. For an analysis of games with imperfect or incomplete information, see, e.g., Gintis [2000]; Myerson [1997]; Osborne and Rubinstein [1994]; Osborne [2004]
• for any vertex \( v \in V \setminus \{v_0\} \), there exists a unique vertex \( w \in V \) and a unique action \( a \in A(w) \) such that \( \text{next}(w,a) = v \);

• for every agent \( i \in \text{Agt} \), the utility function \( U_i : \text{EndV} \to \mathbb{R} \) defines \( i \)'s preferences over the set of terminal nodes. Note that \( U_i \) represents cardinal utilities that can be determined as shown in Section 2.1.

Definition 2.5 specifies a particular type of sequential games, which are with perfect information. This characteristics simply means that, at any moment of the game, all players have a full observability over each other's choices. In other words, a player's only uncertainty in such a game is about other players' choices.

As an illustration of this formalization, one can define the binary ultimatum game from Figure 2.2 as \( \text{EG} = \langle \text{Agt}, V, \text{EndV}, Q, A, \text{next}, \{U_i|i \in \text{Agt}\} \rangle \), where:

• \( \text{Agt} = \{\text{Alice, Bob}\} \);

• \( V = \{v_0, v_1, v_2, v_3, v_4, v_5, v_6\} \);

• \( \text{EndV} = \{v_3, v_4, v_5, v_6\} \);

• \( Q(v_0) = \text{Alice}, Q(v_1) = Q(v_2) = \text{Bob} \);

• \( A(v_0) = \{C_a, D_a\}, A(v_1) = A(v_2) = \{C_b, D_b\} \);

• \( \text{next}(v_0, C_a) = v_1, \text{next}(v_0, D_a) = v_2, \text{next}(v_1, C_b) = v_3, \text{next}(v_1, D_b) = v_4, \text{next}(v_2, C_b) = v_5, \text{next}(v_2, D_b) = v_6 \);

• \( U_{\text{Alice}}(v_3) = U_{\text{Bob}}(v_3) = 5, U_{\text{Alice}}(v_5) = 8, U_{\text{Bob}}(v_5) = 2, U_{\text{Alice}}(v) = U_{\text{Bob}}(v) = 0 \) for every \( v \in \{v_4, v_6\} \).

Following this definition of an extensive game, one needs to define the concept of a strategy in order to be able to determine Nash equilibria. For every agent \( i \in \text{Agt} \), a pure strategy simply specifies an action at every decision node for \( i \). This formally leads to Definition 2.6.

**Definition 2.6 (Pure strategies)** Given an extensive game \( \text{EG} = \langle \text{Agt}, V, \text{EndV}, Q, A, \text{next}, \{U_i|i \in \text{Agt}\} \rangle \), and the set of nodes \( H_J = \{v \in V|\text{EndV}|Q(v) \in J\} \) where any agent \( i \in J \) moves, a pure joint strategy for \( J \) in \( \text{EG} \) is a total function \( s_J \) on \( H_J \) such that, for every \( v \in H_J \), \( s_J(v) \in A(v) \).

For every non-empty group of agents \( J \subseteq \text{Agt} \), let \( S_J \) denote the set of joint strategies for the group \( J \) such that \( S_J = \prod_{i \in J} S_i \). In this case, for notational convenience, we write \( S \) instead of \( S_{\text{Agt}} \), which represents the set of strategy profiles.
(i.e., every \( s \in S \) specifies an action for every vertex of the game), \( S_i \) instead of \( S_{\{i\}} \), and \( S_{-i} \) instead of \( S_{\text{Agt}\setminus\{i\}} \).

Furthermore, let \( P_s \subseteq V \) define the set of vertices that constitutes the actual path predicted by a given strategy \( s \in S \) such that:

- there exists a unique initial vertex \( v_0 \in P_s \), that is, for every vertex \( w \in V \) and every \( a \in A(w) \), \( \text{next}(w,a) \neq v_0 \);
- there exists a unique terminal vertex \( v_n \in P_s \), that is, \( v_n \in \text{EndV} \);
- for every \( v \in P_s \setminus \{v_0,v_n\} \), \( v \) is the successor of another node from \( P_s \), that is, there exists a unique vertex \( w \in P_s \setminus \{v_n\} \) and an agent \( i = \Omega(w) \) such that \( \text{next}(w,s_i(w)) = v \).

Based on the above definition of a pure strategy as well as the previous concept of an actual path, one can express any extensive game in terms of a strategic game, as shown through Definition 2.7.

\textbf{Definition 2.7 (Strategic form game)} Given an extensive game \( EG = \langle \text{Agt}, V, \text{EndV}, \Omega, A, \text{next}, \{U_i|i \in \text{Agt}\} \rangle \), the strategic form of \( EG \) is the strategic game \( G = \langle \text{Agt}, \{S_i|i \in \text{Agt}\}, \{U_i'|i \in \text{Agt}\} \rangle \) where:

- for every agent \( i \in \text{Agt} \), every strategy \( s_i \in S_i \) is defined according to Definition 2.6.
- for every agent \( i \in \text{Agt} \) and every strategy profile \( s \in S \), there exists a unique vertex \( v \in P_s \cap \text{EndV} \) such that \( U_i'(s) = U_i(v) \).

As a result, determining a Nash equilibrium in an extensive game \( EG \) simply corresponds to finding a Nash equilibrium in the strategic form of \( EG \) according to Definitions 2.7 and 2.2 (or Definition 2.4 if considering mixed strategies).

As an illustration of Definition 2.7, let us consider again the binary ultimatum game presented earlier. In fact, it appears that the extensive game depicted in Figure 2.2 can alternatively be represented in terms of the strategic game from Figure 2.3 where the players’ actions are considered as strategies.

In Figure 2.3, while Alice’s set of strategies simply matches her set of actions at vertex \( v_0 \) in the original game (i.e, \( S_{\text{Alice}} = A(v_0) \)), it is slightly more complex to define Bob’s. Indeed, a strategy for Bob consists of a combination of his actions at each of his decision node. More specifically, a strategy \( (C_b,D_b) \) simply specifies that bob would play \( C_b \) if vertex \( v_1 \) is reached, and alternatively, he would play \( D_b \) if vertex \( v_2 \) is reached. Note that this example clearly illustrates the counterfactual aspect of strategies in such sequential games: a strategy does not only indicate
what an agent will actually do, it also expresses what that agent would do at every decision node, including those that may never be reached.

Following such a transformation in a strategic game, it then becomes straightforward to determine the set of Nash equilibria through Definition 2.2 (or Definition 2.4 if considering mixed strategies). For example, applying this principle to the strategic form of the ultimatum game depicted in Figure 2.3 yields three different Nash equilibria: \((C_a, (C_b, C_b))\), \((C_a, (C_b, D_b))\), and \((D_a, (D_b, C_b))\). However, one should note that, according to our initial basic analysis of the game in Figure 2.2, only the solution \((D_a, (C_b, C_b))\) was predicted to be rational. In fact, it appears that the other two solutions rely on some incredible threat made by Bob. As an illustration, let us interpret solution \((C_a, (C_b, D_b))\) as follows: suppose that, at the beginning of the game, the two individuals are allowed to talk to each other, and Bob tells Alice that, if she plays \(D_a\) first, he will then play \(D_b\). If Alice believes this threat, then it becomes rational for her to choose \(C_a\), and consequently, Bob will rationally select \(C_b\). However, it does not make sense for Alice to believe Bob’s threat because it would otherwise imply that Bob will be irrational if Alice first plays \(D_a\) (at \(v_3\), it is indeed never rational for Bob to play \(D_b\)). The same reasoning can also apply to the other solution \((D_a, (D_b, C_b))\). As a result, this example demonstrates that the concept of Nash equilibrium is too general and does not necessarily specify a fully rational solution.

This interpretation therefore suggests a need for an alternative solution concept that allows for such a refinement on the Nash equilibrium that rules out any incredible threat. As a means to introduce such a solution, let us first define the concept of a subgame according to Definition 2.8.

**Definition 2.8 (Subgame)** Given an extensive game with perfect information \(EG = \langle Agt, V, EndV, \Omega, A, next, \{U_i|i \in Agt\} \rangle\), a subgame of \(EG\) is an extensive game with perfect information \(EG' = \langle Agt, V', EndV', \Omega', A', next', \{U'_i|i \in Agt\} \rangle\) such that:

- \(V' \subseteq V\);
- \(EndV' \subseteq EndV\);
• for every vertex \( v \in V' \), \( Q'(v) = Q(v) \);

• for every vertex \( v \in V' \), \( A'(v) = A(v) \);

• for every vertex \( v \in V' \setminus \text{End}V' \) and every action \( a \in A'(v) \), \( \text{next}'(v, a) = \text{next}(v, a) \).

• there exists a unique initial vertex \( v_0 \in V' \setminus \text{End}V' \) such that, for every \( w \in V' \setminus \text{End}V' \) and every \( a \in A'(w) \), \( \text{next}'(w, a) \neq v_0 \);

• for every agent \( i \in \text{Agt} \) and every vertex \( v \in \text{End}V' \), \( U_i'(v) = U_i(v) \).

Following Definition 2.8, one can introduce an alternative equilibrium solution called the subgame perfect equilibrium, which is defined according to Definition 2.9.

**Definition 2.9 (Subgame perfect equilibrium)** Given an extensive game with perfect information \( EG = \langle \text{Agt}, V, \text{End}V, Q, A, \text{next}, \{U_i | i \in \text{Agt}\} \rangle \), a subgame perfect equilibrium is a combination of strategies \( s \in S \) such that \( s \) induces a Nash equilibrium in the normal form of every subgame of \( EG \).

Applying this new concept to the above binary ultimatum game, one can observe that \((D_a, C_b, C_b)\) is the only solution that satisfies Definition 2.9. In other words, the subgame perfect equilibrium can simply be considered as a Nash equilibrium that does not rely on any incredible threat, which leads to state the following theorem.

**Theorem 2.2** Given an extensive game with perfect information \( EG = \langle \text{Agt}, V, \text{End}V, Q, A, \text{next}, \{U_i | i \in \text{Agt}\} \rangle \), if \( \text{Nash} \subseteq S \) defines the set of pure strategy Nash equilibria in the normal form of \( EG \) and \( \text{SPE} \subseteq S \) defines the set of subgame perfect equilibria in \( EG \), then:

\[
\text{SPE} \subseteq \text{Nash}
\]

One should note that there exists an alternative and more intuitive way to compute a subgame perfect equilibrium in some extensive game. Such a method, which relies on some backward induction reasoning, appears to better reflect the way in which the players reason in such sequential situations. More details about this principle will be provided in Chapter 3.

Following the basic principles of game theory that have been introduced in this section, one may argue that the notion of rationality assumed here remains quite ambiguous. Let us therefore clarify what game theory does actually tell us about one’s rationality. The main assumption made by game theory, which directly follows from Section 2.1, is that rationality is strongly based upon the maximization
of one’s own preferences. However, the various analyses in this section clearly indicate that this is not sufficient. In fact, in order to be fully rational in the context of social interactions, it appears that one also needs to consider the preferences of other individuals involved, and assume that they will similarly aim at maximizing those preferences. In other words, game theory introduces a strategic component to the definition of rationality. Such an interpretation is clearly illustrated through the various solution concepts presented here (i.e., the Nash equilibrium and the subgame perfect equilibrium), which reflect the fact that an agent’s rational behavior relies on what everybody would rationally do.

However, despite its high relevance, there is an obvious limitation in such game theoretic thinking: it makes the strong assumption that everything that characterizes a social interaction (i.e., the agents’ actions, strategies, and preferences) is common knowledge among all individuals involved. Moreover, none of the above concepts clearly allows to explicitly capture the more general notion of rationality as an intrinsic property of an individual. One may indeed wonder whether any deviation from these solution concepts should be interpreted as an actual irrational behavior. In order to investigate this issue, we claim that an agent’s knowledge must be taken into consideration as it plays a major role in determining one’s rationality. Such a relevant concept is therefore the object of the following section.

### 2.3 Interactive epistemology

Through Section 2.1 and Section 2.2, we have demonstrated that one’s rationality strongly relies on the preferences of everybody that is involved in a social interaction. However, while it is reasonable to assume that an individual is genuinely aware of his own preferences over outcomes, it is more controversial to assume that he similarly knows the exact preferences of other individuals he is interacting with. And even if all preferences were perfectly known, it is reasonable to assume that an individual may still be uncertain about the way other people will behave. Such an intuitive argument therefore suggests the need for deeply studying the role that knowledge and beliefs play in defining rationality. The concept of epistemic games have been extensively studied in economics in the so-called interactive epistemology area (see e.g., Aumann [1999a]; Aumann and Brandenburger [1995]; Battigalli and Bonanno [1999]; Bonanno [2008]; Brandenburger [1992]; Gintis [2009]). In this kind of games, players decide what to do according to some general principles of rationality while being uncertain about several aspects of the interaction such as other agents’ choices, or other agents’ preferences. More precisely, in Aumann [1995], Aumann elegantly introduces the notion of epistemic rationality, which states that an individual is rational whenever he does not knowingly play a strategy that yields him less than he could have gotten with a different
strategy. The epistemic reference in this definition is crucial as it implies that an individual may actually be rational to choose an action that will not necessarily result in the most optimal option for him (e.g., he may be uncertain about other individuals’ moves or preferences). More interestingly, the role of knowledge also appears to go beyond the definition of rationality itself. In fact, the overly idealistic assumption that is made in classical game theory is that this concept of rationality is commonly known among all rational agents. Many existing works have already investigated this important issue (see e.g., Aumann [1995]; Aumann and Brandenburger [1995]; Battigalli and Bonanno [1999]; Brandenburger [1992]) through analysing the necessary and/or sufficient epistemic conditions of various equilibrium notions (e.g., the Nash equilibrium, the subgame perfect equilibrium). Throughout this section, we therefore present two relevant approaches that allow to formalize such reasoning about knowledge: Aumann’s set-theoretic model, and modal logic.

2.3.1 Aumann’s common knowledge

In Aumann [1999a], Aumann introduces a simple way to formalize the notion of interactive knowledge. We here describe the corresponding set-theoretic representation that leads to a mathematical characterization of common knowledge. For this purpose, let us define Aumann’s epistemic model as follows.

Definition 2.10 (Aumann’s epistemic model) Aumann’s epistemic model is a tuple $A = \langle \text{Agt}, \Omega, \xi, \{I_i|i \in \text{Agt}\} \rangle$ where:

- $\text{Agt}$ is a non-empty set of agents;
- $\Omega$ is a non-empty set of states of the world;
- $\xi$ is a non-empty set of outcomes $E \subseteq \Omega$, each of which consists of a subset of states;
- every $I_i$ is a total function $I_i : \Omega \rightarrow 2^{\Omega}$ that represents the set of states that agent $i$ cannot distinguish from $\omega \in \Omega$ such that, $\omega \in I_i(\omega)$, and if $\omega' \in I_i(\omega)$, $I_i(\omega') = I_i(\omega)$.

Note that an outcome $E$ is defined as a subset of states of the world. For example, assuming that $E$ describes the outcome that “it will rain tomorrow”, $E$ consists of all possible states from $\Omega$ at which it will rain tomorrow.

Furthermore, for any two states $\omega, \omega' \in \Omega$ and any agent $i \in \text{Agt}$, $I_i(\omega)$ and $I_i(\omega')$ are either disjoint (i.e., $I_i(\omega) \cap I_i(\omega') = \emptyset$) or identical (i.e., $I_i(\omega) = I_i(\omega')$). This implies that every function $I_i$ simply forms an information partition of $\Omega$, that is, $\cup_{\omega \in \Omega} I_i(\omega) = \Omega$. 

24
Then, for every agent $i \in Agt$, one can construct a total function $K_i : \xi \to \xi$ such that $K_i(E)$ defines the set of all states $\omega \in \Omega$ at which $i$ knows that $E$ contains $\omega$, that is:

$$K_i(E) = \{\omega \in \Omega | I_i(\omega) \subseteq E\}$$

In other words, $K_i(E)$ simply reads “agent $i$ knows that outcome $E$ obtains”. As a result of this definition, one can notice that the knowledge operator $K_i$ has the properties depicted in Table 2.1 (for all outcomes $E, F \subseteq \Omega$).

According to Property (2.1), an agent can only know something if it is true. Property (2.2) says that, if an outcome $E$ entails another outcome $F$, then knowing $E$ entails knowing $F$. Property (2.3) expresses that, if an agent does not know something, then he knows that he does not know it. This principle is often called negative introspection. Similarly, Property (2.4) defines what is often called positive introspection, that is, if an agent knows something, then he knows that he knows it. Finally, according to Property (2.5), an agent knows two things if and only if he knows each of them independently.

As a result, the knowledge operators $K_i$ for every agent $i \in Agt$ allows to express the concept of common knowledge. An outcome $E$ is common knowledge among all members of $Agt$ if and only if all know $E$, all know that all know $E$, all know that all know that all know $E$, and so on ad infinitum. Formally, let $CK(E)$ denote that outcome $E$ is common knowledge among all members of $Agt$: $CK(E) = K^1(E) \cap K^2(E) \cap K^3(E) \cap \ldots$

where $K^1(E) = \cap_{i \in Agt} K_i(E)$ and, for any $m > 0$, $K^{m+1}(E) = K^1K^m(E)$.

However, while Aumann states that such a semantic representation is convenient to use (it is indeed based on a simple framework, as shown through Definition 2.10), he agrees that it is also not so straightforward and involves some non-trivial issues (Aumann [1999a]). More specifically, he points out some relevant questions that arise from this strictly semantic approach: do the agents know about the

\[(2.1) \quad K_i(E) \subseteq E\]
\[(2.2) \quad E \subseteq F \text{ implies that } K_i(E) \subseteq K_i(F)\]
\[(2.3) \quad \neg K_i(E) \subseteq K_i(\neg K_i(E))\]
\[(2.4) \quad K_i(E) \subseteq K_i(K_i(E))\]
\[(2.5) \quad K_i(E \cap F) = K_i(E) \cap K_i(F)\]

Table 2.1: Some properties of operator $K_i$
model itself? In other words, do they know about the information partitions of the set of states (i.e., the functions $I_i$)? What explicitly describes a state of the world? In fact, a complete description of a state of the world $\omega$ would include a list of all the states that cannot be distinguished from $\omega$ by any particular agent. In other words, every state should describe every agent’s uncertainty at that state. More generally, Aumann argues in Aumann [1987, 1999a] that “such a description of the $\omega$’s involves no ‘real’ knowledge; it is only a kind of code book or dictionary”. In other words, one’s knowledge simply depends on what is the true state of the world. As a result, while an agent $i$ may be ignorant about what another agent $j$ knows, $i$ cannot be ignorant about $j$’s information function $I_j$.

In order to allow for a more explicit description of knowledge and clarify all such ambiguities, the most convincing solution lies in a syntactic approach, which appears to be more intuitive while expressing any such concepts with transparency. Therefore, as this analysis clearly suggests the need for a correspondence between both a semantic and a syntactic representation, we claim, through the next section, that modal logic represents the ideal tool to formally characterize knowledge in interactive situations.

### 2.3.2 Epistemic modal logic

Modal logic is similar to traditional propositional logic with the addition of extra operators that can be used to express modalities of truth, including possibility and necessity. Epistemic modal logic then represents a subclass of such a modal logic that allows to reason about knowledge (Blackburn et al. [2002]; Chellas [1980]; Hintikka [1962]; Hughes and Cresswell [1968]; van der Hoek and Pauly [2006]).

In order to justify the need for such a logical formalism, let us consider the following famous puzzle. Two children, say Alice and Bob, come back from the garden, both with mud on their forehead. Their father looks at them and says:

“at least one of you has mud on his/her forehead”

Then he asks:

“those who know whether they are dirty, step forward!”

In this case, nobody steps forward. So the father asks again:

“those who know whether they are dirty, step forward!”

This time, both Alice and Bob simultaneously step forward and claim to know that they are dirty.
As a means to explicitly interpret the epistemic reasoning that is involved in this scenario (which we will refer to as the muddy children puzzle), let us construct a basic epistemic logic as follows.

First, we consider a set of atomic propositions $Atm$ and a non-empty finite set of agents $Agt$. The language of such an epistemic logic (Halpern and Moses [1992]) is then given by the following BNF (Backus-Naur Form) grammar:

$$\varphi ::= p \mid \bot \mid \neg \varphi \mid \varphi \lor \varphi \mid [K_i] \varphi$$

where $p$ ranges over $Atm$ and $i$ ranges over $Agt$. The classical Boolean connectives $\top$, $\land$, $\rightarrow$ and $\leftrightarrow$ can then be abbreviated as follows:

- $\top \overset{\text{def}}{=} \neg \bot$
- $\varphi \land \psi \overset{\text{def}}{=} \neg (\neg \varphi \lor \neg \psi)$
- $\varphi \rightarrow \psi \overset{\text{def}}{=} \neg \varphi \lor \psi$
- $\varphi \leftrightarrow \psi \overset{\text{def}}{=} (\varphi \rightarrow \psi) \land (\psi \rightarrow \varphi)$

The intuitive meaning of $[K_i] \varphi$ is “agent $i$ knows that $\varphi$ is true”. $[K_i]$ is called a modal operator. Moreover, one can add the following abbreviation:

$$\langle K_i \rangle \varphi \overset{\text{def}}{=} \neg [K_i] \neg \varphi$$

In this case, $\langle K_i \rangle \varphi$ reads “it is possible for $i$ that $\varphi$ is true”.

As a consequence, a formula in this logic simply consists of any combination of the elements from the above language. As an example, let two propositions $muddyAlice$ and $muddyBob$ express the respective facts that “Alice has mud on her forehead” and “Bob has mud on his forehead”. In this case, the formula $[K_{Alice}]muddyBob \land \langle K_{Alice} \rangle muddyAlice$ means that “Alice knows that Bob has mud on his forehead, and it is possible for her that she also has mud on her forehead”.

One then needs to specify the conditions under which a given formula, like the one above, is true or false. For this purpose, we introduce a model that gives a formal meaning to all operators of the above language.

The models that we are considering here rely on a possible worlds semantics, as it was introduced by Kripke (Kripke [1963]). The basic idea of such a model is that it considers a set of possible worlds, each of which describing a particular state (as in Aumann’s model from Section 2.3.1), and only one of those is reported to be the actual world. An agent’s epistemic state is then defined in each world $w$ as the set of worlds that cannot be distinguished from $w$.

Formally, such a model is defined as follows.
Definition 2.11 (Epistemic model) An epistemic model is a tuple \( M = \langle W, \{ \mathcal{E}_i | i \in \text{Agt} \}, \pi \rangle \) where:

- \( W \) is a non-empty set of possible worlds;
- every \( \mathcal{E}_i \) is an equivalence relation on \( W \);
- \( \pi : \text{Atm} \rightarrow 2^W \) is a valuation function on worlds.

In Definition 2.11, such a model \( M \) is sometimes referred to as a type of Kripke structure. Given a model \( M = \langle W, \{ \mathcal{E}_i | i \in \text{Agt} \}, \pi \rangle \), the pair \( F = \langle W, \{ \mathcal{E}_i | i \in \text{Agt} \} \rangle \) is often called a frame (Blackburn et al. [2002]), while \( \pi \) is called a valuation.

More specifically, for every agent \( i \in \text{Agt} \), all relations \( \mathcal{E}_i \) characterize \( i \)'s knowledge in \( M \). In this case, note that every relation \( \mathcal{E}_i \) is:

- Reflexive: for every \( w \in W \), \( w \mathcal{E}_i w \);
- Symmetric: for every \( v, w \in W \), \( v \mathcal{E}_i w \) implies \( w \mathcal{E}_i v \);
- Transitive: for every \( v, w, q \in W \), \( v \mathcal{E}_i w \) and \( w \mathcal{E}_i q \) together imply \( v \mathcal{E}_i q \);
- Euclidean: for every \( v, w, q \in W \), \( v \mathcal{E}_i w \) and \( v \mathcal{E}_i q \) together imply \( w \mathcal{E}_i q \).

Given two worlds \( w, v \in W \), \( w \mathcal{E}_i v \) means that agent \( i \) cannot distinguish world \( v \) from world \( w \). We alternatively write \( \mathcal{E}_i(w) \), describing the set of worlds that \( i \) cannot distinguish from \( w \). One should note that these relations \( \mathcal{E}_i \) are analogous to Aumann’s information partition \( I_i \) as it was introduced in Definition 2.10 from Section 2.3.1.

As an illustration, let us represent a Kripke model that characterize the above muddy children initial situation. Therefore, we suppose that \( \text{Agt} = \{ \text{Alice}, \text{Bob} \} \) and \( \text{Atm} = \{ \text{muddyAlice}, \text{muddyBob} \} \). In this case, we can construct the model \( M \) corresponding to the above muddy children puzzle as follows:

- \( W = \{ w_1, w_2, w_3, w_4 \} \);
- \( \mathcal{E}_{\text{Alice}}(w_1) = \mathcal{E}_{\text{Alice}}(w_2) = \{ w_1, w_2 \}, \mathcal{E}_{\text{Alice}}(w_3) = \mathcal{E}_{\text{Alice}}(w_4) = \{ w_3, w_4 \} \), \( \mathcal{E}_{\text{Bob}}(w_1) = \mathcal{E}_{\text{Bob}}(w_3) = \{ w_1, w_3 \}, \mathcal{E}_{\text{Bob}}(w_2) = \mathcal{E}_{\text{Bob}}(w_4) = \{ w_2, w_4 \} \);
- \( \pi(\text{muddyAlice}) = \{ w_1, w_3 \}, \pi(\text{muddyBob}) = \{ w_1, w_2 \} \).

Note that any such model can easily be represented as a graph where the vertices correspond to the worlds, and the edges are defined through every relation \( \mathcal{E}_i \). Figure 2.4 depicts such a graphical representation of the above muddy children model \( M \).

In order to define how to evaluate modal formulas over such a Kripke model, one then needs to give a formal meaning to every basic operator of the language, which can be done through the following truth conditions.
Definition 2.12 (Truth conditions) Truth of a formula in a model $M$ at a given world $w$ is defined as follows:

- $M, w \models p$ iff $w \in \pi(p)$;
- $M, w \models \neg \varphi$ iff $M, w \not\models \varphi$;
- $M, w \models \varphi \lor \psi$ iff $M, w \models \varphi$ or $M, w \models \psi$;
- $M, w \models [K_i] \varphi$ iff $M, w' \models \varphi$ for all $w'$ such that $w \not\in \mathcal{E}_i w'$.

According to Definition 2.12, $M, w \models \varphi$ reads “formula $\varphi$ is true at world $w$ in model $M$”. In this case, we say that a formula $\varphi$ is true in a model $M$ (noted $M \models \varphi$) if and only if $M, w \models \varphi$ for every world $w$ in $W$. Furthermore, $\varphi$ is valid (noted $\models \varphi$) if and only if $\varphi$ is true in all possible Kripke models. Then we further say that $\varphi$ is satisfiable iff $\neg \varphi$ is not valid.

Regarding the knowledge operator $[K_i]$, the formula $[K_i] \varphi$ is true at a world $w$ in the model $M$ if and only if $\varphi$ holds in all worlds that are considered possible for agent $i$. In this case, the higher the number of such worlds, the more $i$ is ignorant at $w$. On the other hand, if $i$ knows everything at $w$, then it means that $w$ is the only possible world for $i$, that is, $\mathcal{E}_i(w) = \{w\}$.

Following this interpretation of knowledge, one should note the distinction that such a logical framework makes with Aumann’s epistemic model from Section 2.3.1. In fact, it is clearly suggested here that one’s knowledge is not to be evaluated in
the whole model, but instead in what appears to be the true state of the world: as a consequence, one’s knowledge may reasonably be different depending on what the actual world is. Such a transparent and straightforward representation therefore allows to resolve all ambiguities that were previously raised through Aumann’s model (see Section 2.3.1).

However, while the previous epistemic logic clearly provides the combination of a semantic representation with a syntactic formalism, one still needs to explicitly relate the two approaches and formalize their exact correspondence. This step can be done through an axiomatization of the logic. The formulas from Table 2.2 correspond to axioms related to every knowledge operator $[K_i]$ (for every agent $i \in Agt$), which follow from the above properties of the equivalence relation $\mathcal{E}_i$: these formulas are therefore necessarily true at all worlds in all Kripke models.

In Table 2.2, Axiom $K$, which is usually called the distribution axiom, expresses understanding of the material implication, that is, if some agent $i$ knows that $\varphi$ and $i$ knows that $\varphi$ implies $\psi$, then $i$ also knows $\psi$. Axiom $T$, which is called the truth axiom, indicates that an agent can only know facts. Furthermore, Axiom 4 expresses the agents’ positive introspection through the fact that if an agent knows something, then he knows that he knows it. Similarly, negative introspection is expressed through Axiom 5, which states that if some agent $i$ does not know something, then he knows that he does not know it.

In this case, the current epistemic logic is completely axiomatized by all principles of classical propositional logic and the set of axioms from Table 2.2. In this case, we write $\vdash \varphi$ to mean that $\varphi$ can be derived by means of the principles given in Table 2.2 and those of classical propositional logic. The resulting logic is called $S5_n$ where $n$ denotes the number of agents in $Agt$ (i.e., $n = |Agt|$). Traditionally, $S5$ is the logic with only one knowledge operator associated with one equivalence relation, and so $S5_n$ simply corresponds to the fusion of $n$ logic $S5$.

Moreover, one should note that such a $S5_n$ logic also consists of logical laws that allow to draw conclusions from premises. Examples of such inference rules are introduced through Table 2.3.

| $K$ | $[K_i](\varphi \to \psi) \to ([K_i]\varphi \to [K_i]\psi)$ |
| $T$ | $[K_i]\varphi \to \varphi$ |
| 4 | $[K_i]\varphi \to [K_i][K_i]\varphi$ |
| 5 | $\neg[K_i]\varphi \to [K_i]\neg[K_i]\varphi$ |

Table 2.2: Axiom schemas for every modal operator $K_i$ (S5)
\[\varphi, \varphi \rightarrow \psi \quad \Rightarrow \quad \psi\]

(Necessitation)
\[\varphi \quad \Rightarrow \quad [K_i]\varphi\]

(Monotony)
\[\varphi \rightarrow \psi \quad \Rightarrow \quad [K_i]\varphi \rightarrow [K_i]\psi\]

Table 2.3: Inference rules

In Table 2.3, **Modus Ponens** expresses that if the formulas \(\varphi\) and \(\varphi \rightarrow \psi\) are true in the model (i.e., \(M \models \varphi \land (\varphi \rightarrow \psi)\)), then so is \(\psi\) (i.e., \(M \models \psi\)). Similarly, **Necessitation** states that if a formula \(\varphi\) is true in the model, then everybody knows that \(\varphi\) is true. **Monotony** further indicates that if an implication is true in the model, then everybody knows about it.

Note that the axioms from Table 2.2 along with the rules of inference from Table 2.3 correspond to the properties of Aumann’s knowledge functions as they were introduced in Section 2.3.1 (see Table 2.1).

Furthermore, it is worth noting that Aumann’s concept of common knowledge, as it was introduced semantically in Section 2.3.1, can also be expressed in the language of this epistemic logic. In fact, let us define \([EK]\varphi\) as an abbreviation of \(\land_{i \in \text{Agt}}[K_i]\varphi\), i.e. every agent knows \(\varphi\) (if \(C = \emptyset\) then \([EK]\varphi\) is equivalent to \(\top\)). Then we define by induction \([EK^k]\varphi\) for every natural number \(k \in \mathbb{N}\):

\[\begin{align*}
[EK^0]\varphi & \overset{\text{def}}{=} \varphi \\
\text{and for all } k \geq 1, \\
[EK^k]\varphi & \overset{\text{def}}{=} [EK]([EK^{k-1}]\varphi)
\end{align*}\]

Similarly, we define for all natural numbers \(n \in \mathbb{N}\):

\[\begin{align*}
[CK^0]\varphi & \overset{\text{def}}{=} \varphi \\
\text{and for all } n \geq 1, \\
[CK^n]\varphi & \overset{\text{def}}{=} \land_{1 \leq k \leq n} [EK^k]\varphi
\end{align*}\]

\([CK^n]\varphi\) therefore expresses common knowledge that \(\varphi\) up to \(n\) iterations, i.e. everyone knows \(\varphi\), everyone knows that everyone knows \(\varphi\), and so on until level \(n\).

As a means to illustrate the syntactic power of such an epistemic logic, let us go back to our muddy children puzzle. According to the model \(M\) depicted in Figure
2.4, which represents both children’s background knowledge, we assume that the actual world corresponds to that where both Alice and Bob are muddy (i.e., world \( w_1 \)). In this case, one can express the following facts about Alice that are true at \( w_1 \) (similar statements can be made for Bob):

- “Alice knows that Bob is muddy”:
  \[
  M, w_1 \models [K_{Alice}]\text{muddyBob}
  \]

- “Alice knows that Bob knows whether she is muddy or not”:
  \[
  M, w_1 \models [K_{Alice}]( [K_{Bob}]\text{muddyAlice} \lor [K_{Bob}]\neg\text{muddyAlice})
  \]

- “Alice knows that Bob does not know that she is muddy if and only if he knows that she is not muddy”:
  \[
  M, w_1 \models [K_{Alice}](\neg[K_{Bob}]\text{muddyAlice} \iff [K_{Bob}]\neg\text{muddyAlice})
  \]

Furthermore, let us now concentrate on what Alice gets to learn from the father’s public announcement that “at least one of the children is muddy”, which becomes common knowledge among both children. In this case, the following facts become true at the actual world \( w_1 \) in the updated model \( M' \) depicted in Figure 2.5:

- “it is common knowledge among Alice and Bob that at least one of them is muddy”:
  \[
  M', w_1 \models [CK^n](\text{muddyAlice} \lor \text{muddyBob}) \text{ for any } n \in \mathbb{N}
  \]

- “Alice knows that Bob knows that one of them is muddy”:
  \[
  M', w_1 \models [K_{Alice}][K_{Bob}](\text{muddyAlice} \lor \text{muddyBob})
  \]

- “Alice knows that, if Bob does not know that he is muddy, then she is muddy”:
  \[
  M', w_1 \models [K_{Alice}](\neg[K_{Bob}]\text{muddyBob} \rightarrow \text{muddyAlice})
  \]

As shown in Figure 2.5, the corresponding semantic interpretation of these statements is that the father’s announcement simply rules out the world \( w_4 \) as a possibility to be the actual world (in \( w_4 \), no children is muddy, which contradicts the father’s announcement). In other words, Alice is now only uncertain about whether the actual world is \( w_1 \) or \( w_2 \). Similarly, Bob is only uncertain about whether the actual world is \( w_1 \) or \( w_3 \).

Finally, let us look at what Alice learns from the fact that Bob does not step forward after the father’s first announcement. The following formulas then hold at \( w_1 \) in the updated model \( M'' \) depicted in Figure 2.6:
• “Alice knows that Bob does not know that he is muddy”:

\[ M'', w_1 \models [K_{Alice}] (\neg [K_{Bob}] \text{muddy Bob}) \]

• “Alice knows that she is muddy”:

\[ M'', w_1 \models [K_{Alice}] \text{muddy Alice} \]

As shown in Figure 2.6, the corresponding interpretation that follows is that both children know exactly that \( w_1 \) is the actual world.

However, the above logic may also be subject to some criticism regarding its axiom system and the various rules of inference. In fact, one may argue from Table 2.2 that the truth axiom \( T \) is very strong and therefore not so realistic when modeling human reasoning. An agent may indeed believe something to be true even though it is not. In order to allow for such a weaker notion of knowledge,
Table 2.4: Alternative axiom schemas for the modal operator $K_i$ (KD45)

One may redefine the semantics of operator $[K_i]$ so that it replaces the truth axiom $\mathbf{T}$ with a seriality axiom $\mathbf{D}$. The resulting axiom schema for every $[K_i]$ is depicted in Table 2.4 and define what is usually called a doxastic logic (note that in such a logic, the belief operator $[B_i]$ is generally used instead of $[K_i]$).

In order to obtain the axiom schema from Table 2.4, one must also redefine the epistemic relations $E_i$ for every agent $i \in \text{Agt}$: $E_i$ must indeed be transitive, Euclidean, and serial (i.e., for every $w \in W$, there exists some $v \in W$ such that $w E_i v$). Furthermore, it is worth noting that other similar criticisms can still apply to such a logic, especially concerning the axioms of introspection ($4$ and $5$).

More generally, another problem that can arise from such a system concern logical omniscience. In fact, it appears that the previous axioms along with the rules of inference listed in Table 2.3 can lead to some unrealistic assumptions about the reasoning power of an agent. In fact, according to the above logic, if I know all the axioms and inference rules of Peano arithmetic, then I know whether Fermat’s last theorem is true or false. The limitations of cognitive abilities clearly point out a weakness of the above logical formalism for modeling human reasoning.

However, despite the undeniable relevance of this remark, the epistemic logic presented here appears to be extremely useful in the field of game theory as it suffices to investigate the type of reasoning process that may lead an agent to follow a particular behavior. More specifically, not only this formalism clearly allows to capture the epistemic characterization of rationality as it is introduced by Aumann in Aumann [1995], it also allows to identify the necessary and sufficient conditions that lead rational agents to play according to the game-theoretic predictions.

Following the mathematical approaches of various economists (see, e.g., Aumann [1995]; Aumann and Brandenburger [1995]; Brandenburger [1992]) to investigate these issues, this argument justifies the more recent growing interest for combining modal logic with game theory in order to provide an in-depth analysis of epistemic games both in strategic form and in extensive form (see, e.g., Battigalli and Bonanno [1999]; Bonanno [2002]; Lorini [2010]; Lorini and Schwarzentzuber [2010]; Van Benthem [2003]; van Benthem [2007]). In fact, the various logical
frameworks that are proposed in those works clearly allow to reason about actions, preferences, and knowledge. However, there also appears to be another important aspect that has been much less studied in such games: temporal reasoning. Obviously, the type of social interactions that we are referring to here consists of sequential games. As shown by the distinction between strategic games and extensive games in Section 2.2, considering the temporal component of such interactions leads to dramatically more complex representations. Indeed, while a strategy in a strategic games simply consists of a type of action (see Section 2.2.1), a strategy in an extensive form game expresses not only the sequence of actions that will occur next, but also the actions that would occur in every vertex of the game (see Section 2.2.2 for details). Although we demonstrated in Section 2.2.2 that any extensive game could be expressed in terms of a strategic game, it is clear that the temporal factor of such games can have a non-negligible effect on the agents’ way of thinking: an agent may then learn as he advances through the game, based on what has happened before. Similarly, one may forget about what happened before, or one may even think about what could have happened had another course of action taken place.

Yet, despite the high relevance of such epistemic games in extensive form, there exists no logic that has been shown to be sufficiently general to reason about actions, preferences, knowledge, and time. While it is shown in van Benthem et al. [2011] and van Benthem et al. [2006] that reasoning about actions only is sufficient to compute solution concepts like the backward induction, such game logics cannot express the notion of substantive rationality as in Aumann’s definition from Aumann [1995], which fully considers the temporal aspect of the concept of strategy (i.e., current and future choices, as well as counterfactual choices). In de Bruin [2004], a logic, which enables to reason about the epistemic aspects of extensive games, can deal with several game-theoretic concepts such as knowledge, rationality and backward induction. Nevertheless, all these notions are atomic propositions of the logic managed by an ad hoc axiomatization. Moreover, the logical approach to extensive form games proposed in de Bruin [2004] is purely syntactic: no model-theoretic analysis of extensive form games is proposed.

It is therefore our attempt to fill this gap in Chapter 3 by introducing a logical formalism with a corresponding formal semantics for extensive form games, which allows:

• to express in the object language time-based solution concepts like backward induction,

• to derive syntactically the epistemic and rationality conditions on which such solution concepts are based.

We will also show, in Chapter 3, that such a detailed epistemic analysis further
allows to clearly identify the bounds of rationality. In fact, in many interactive situations, it appears that most (if not all) actions available at a given moment can be performed rationally by some individual with limited knowledge. In other words, any such choice can be justified by some sound and rational arguments. However, while such an interpretation can allow to explain some observed deviation from a game-theoretic prediction, one should note that it may not always suffice. We investigate this issue through the next section.

2.4 Towards social rationality

Although we have already provided a rather precise definition of some crucial aspects of rationality through the previous sections of this chapter, there still remains one major component that also needs to be considered. This component relates to the fact that human beings are social creatures, living their life through a high level of interactions with one another. One may then argue that such a social feature is already taken into account by classical game theory. In fact, we have shown, through Section 2.2, that one’s behavior is not only driven by one’s own preferences, it also largely relies on the preferences of other individuals involved in the interaction. However, according to game theory, an agent’s motivation is assumed to be purely individualistic, driven by the ultimate goal of maximizing one’s own profit by taking advantage of others (whatever it may cost them). In other words, while game theory clearly allows to specify strategic behavior, it does not say anything about actual social behavior, which consists in making choices that are oriented towards the welfare of the group, rather than the individual. Yet the plausibility of such other-regarding behavior is clearly justified by the existence of social attitudes such as altruism, fairness, justice, reciprocity, and equity, which we all constantly witness in our everyday lives.

As a consequence, the high relevance of such properties of social behavior suggests an extension of game theory that relies on empirical measurements. The purpose of this section is therefore to introduce such an experimental methodology, and demonstrate how it allows to more accurately model human rational behavior in the context of social interactions.

2.4.1 Experimental game theory

Running laboratory experiments has only recently become a common practice in the field of economics. Following the work of Chamberlin who conducted the first market experiment (Chamberlin [1948]), many have similarly obtained important results using this approach, either to support traditional economic theory or to develop more accurate models of decision making (see, e.g., Berg et al. [1995];
Bolton and Ockenfels [2000]; Charness and Rabin [2002]; Fehr and Schmidt [1999]; Kahneman and Tversky [1979]; Levine [1998]; Roth [1995]; Smith [1962]). One of the main advantages of all these empirical studies is that they come as a complement of the important theoretical work (see the previous sections). This approach has led to the development of experimental economics, which can be distinguished from experimental social psychology that ignores the existence of an analytical tool as powerful as game theory to model observed social behavior.

Moreover, another major interest of using game theoretic tools concerns the design of the experiment itself. In fact, the abstract representation of social interactions, as proposed by game theory, clearly allows to determine what situation would be best suited to answer a particular research question. For example, one can distinguish an important subclass of games, which characterize situations that hold “zero-sum” properties. A zero-sum game basically corresponds to a social interaction in which any participant’s gain (or loss) of utility is exactly balanced by the losses (or gains) of the other participants. In other words, if the total gains of the players are added up, and the total losses are subtracted, they will sum to zero. A very intuitive example of such a zero-sum situation is the game of chess, in which both players repeatedly interact with each other so that only one of them eventually wins, thereby implying that the other loses. In the context of such a game, one can state that the theory of rationality presented through the previous sections then suffices to accurately explain the corresponding behavior elicited by human agents. In this case, although one may reasonably argue that people do not often play chess very optimally, one may justify it by their cognitive limitations, which simply prevents them from fully internalizing the extremely complex situation.

Historically, the original motivation of studying zero-sum games, which later led to the development of non-cooperative game theory itself (Von Neumann and Morgenstern [1944]), was based on the assumption that many types of real-life social interactions shared this zero-sum property (e.g., business transactions). However, many studies have since indicated that such a strong and pessimistic assumption clearly appears to oversimplify reality. As it has been theorized in Wright [2000], society becomes increasingly non-zero-sum as it becomes more interdependent. This argument can also be summarized as follows:

“The more complex societies get and the more complex the networks of interdependence within and beyond community and national borders get, the more people are forced in their own interests to find non-zero-sum solutions. That is, win-win solutions instead of win-lose solutions.” (Bill Clinton, 2000)

As most empirical works in economics involve the study of such non-zero-sum games, let us clarify the reason why they are often considered as a threat to the
concept of rationality.

The first main issue in such games concerns the notion of co-operation, which simply consists of the process of individuals working or acting together. It is indeed straightforward to observe that cooperation can only be involved in situations that carry out some non-zero-sum property (zero-sum games instead necessarily imply the existence of some “losing” individual). In order to illustrate the main problem raised by cooperation, we consider a specific category of simultaneous games (cf. Section 2.2.1) that involve a coordination problem, i.e., a situation where the players have to coordinate with one another in order to reach a profitable outcome. More specifically, suppose a situation where two individuals (Alice and Bob) face the same choice between two options $A$ and $B$. Assuming those players have no way to communicate with each other, if they both select $A$, then they will each receive $5\,\text{€}$, and if they both select $B$, then they will each receive $1\,\text{€}$. However, if both select different options, then they will get nothing (i.e., $0\,\text{€}$). What should each individual do in this case? This very simple scenario, whose payoff matrix is depicted in Figure 2.7, is well known in the economics literature as the Hi-Lo matching game.

$$
\begin{array}{c|cc}
 & A & B \\
\hline
A & (5, 5) & (0, 0) \\
B & (0, 0) & (1, 1) \\
\end{array}
$$

Figure 2.7: Hi-Lo matching game

The nice property of this simple game is that it clearly illustrates a limitation of classical game theory. In fact, as shown in Bacharach [2006]; Colman et al. [2008], when actually playing this simple game, people largely coordinate on the most rewarding outcome for both individuals, that is, they both select $A$. However, even though such an empirical result clearly appears to be intuitive, it cannot be reached through some classical game theoretic reasoning. While considering the game from Figure 2.7, one can indeed distinguish two different Nash equilibria: $(A, A)$ and $(B, B)$. In other words, game theory applied to this scenario simply remains indecisive (i.e., both playing $A$ and playing $B$ can be selected rationally). While predicting the solution $(A, A)$ is not problematic as it corresponds to what people actually do, let us explain why $(B, B)$ is also predicted as a credible outcome. In order for Alice to play $B$, she must believe that Bob will also play $B$. Yet one may wonder how she could form such a belief state. This belief can indeed be justified by her belief that Bob believes that she plays $B$. Similarly, this belief can be
reached as a consequence of believing that Bob believes that she believes that he plays \( B \). This reasoning, which can obviously be repeated indefinitely illustrates the concept of common knowledge or belief, as it was presented in Section 2.3. As a result of this interpretation, the game from Figure 2.7 clearly illustrates the failure of classical game theory to explain how people actually behave the way they do. More precisely, it also demonstrates the limitation of the theory of rationality that has been considered so far. Note that, although some refinements to the Nash equilibrium solution concept have been proposed in the past (see, e.g., the notions of risk dominance and payoff dominance proposed in Harsanyi and Selten [1988], which exclude the \((B, B)\) outcome to be a valid equilibrium solution in the Hi-Lo game from Figure 2.7), those do not allow to explain how individual rationality should be revised in order to reach the predicted behavior, as it is argued in Bacharach [2006]. In order to fill this gap, some theories of bounded rationality have been considered, such as the principle of insufficient reason (see, e.g., Gintis [2003]), which can be described as follows in the context of the above Hi-Lo game: as a best response to believing that the co-player has insufficient reason and will randomize over the set of actions (i.e., select \( A \) and \( B \) with equiprobability), one’s only rational move is therefore to select option \( A \) (under this belief, selecting \( B \) is no longer rational). While it is clear that such a bounded rationality argument indeed suffices to justify the observed behavior in the Hi-Lo game, it also reveals the following weakness: in order to predict outcome \((A, A)\) as the unique solution, this asymmetric model (each player treats the other as unlike himself) requires each agent to wrongly believe in the other’s insufficient reason, which is clearly a strong and unrealistic assumption. The extreme simplicity of this game indeed makes it reasonable to claim that, on the contrary, each player rationally selects \( A \) while believing in each other’s full rationality. As a result, the above Hi-Lo game clearly represents a paradox that standard models of game theory cannot solve, as indicated in Bacharach [2006].

Furthermore, the difficulty for explaining such rational cooperation points out to yet another important issue that can often be found in non-zero-sum games: social dilemmas. Such dilemmas simply consist of situations in which collective interests are in conflict with private interests (again, note that, for obvious reasons, such dilemmas cannot be found in zero-sum games). In many economic studies, the use of games that involve such dilemmas is indeed often preferred, as it offers a credible alternative to the classical prediction of rationality. A well known example of such a game is the prisoner’s dilemma that was introduced in Section 2.2.1 (see Figure 2.1). As shown before, this game indeed yields a unique rational choice that leads to mutual defection. However, mutual cooperation appears to be also appealing as it would then result in the best outcome for the group, and it is a better solution for each individual than mutual defection (i.e., it corresponds to a
“win-win” solution). In fact, the various studies involving this particular game have provided empirical evidence that such a social dilemma has some non-negligible effect on human behavior. For example, it is shown in Shafir and Tversky [1992], that the cooperation rate in such scenarios varies between 30-40%.

Alternatively, such social dilemmas can also be found in sequential games (cf. Section 2.2.2), which are therefore regularly used in experiments to test the robustness of the classical game-theoretic models assuming selfish preferences and equilibrium. In this case, the advantage of the temporal component of such games is manifold: it allows to measure the effects that important psychological factors such as reciprocity, trust and emotions, can have on behavior. As an illustration of this kind of situations, let us consider the following game that again involves two players: first, Alice is given a certain amount of money, say 20 €, and is asked to choose either to fairly share the amount (i.e., 10 € for each player), which ends the game, or to give the full amount away to Bob. In the latter case, Bob receives the amount offered by Alice multiplied by 3 (i.e., $3 \times 20 = 60$ €) and then faces the following dilemma: he has to either fairly share the total amount with Alice (30 € for each player), or to keep the whole amount for himself. In the latter case, Bob therefore earns 60 € while Alice earns nothing and the game ends. The graphical representation of this trust game, as it is usually called in economics (Berg et al. [1995]), is depicted in Figure 2.8.

Applying game theory to the game in Figure 2.8 yields a unique rational argument: if given the chance, Bob’s only rational move is then to play $D$. Knowing that, Alice’s unique best response is therefore to play $D$. However, an alternative interpretation of this game can be the following: by playing $C$, Alice then “trusts” Bob to later perform $C$. As a result, playing $C$ for Bob corresponds to a form of positive reciprocity to Alice’s initial trust. Furthermore, a similar form of negative reciprocity (i.e., punishment) can also be found in a variant of the above game,
which is called the ultimatum game (for an illustration, see Figure 2.2 in Section 2.2.2).

As another interesting property of such games, one should note that they can also allow the elicitation of various types of social emotions. For example, one may argue that Bob’s choice to play $C$ in the above trust game from Figure 2.8 is led by some form of guilt aversion, as suggested in Battigalli and Dufwenberg [2007]; Ellingsen et al. [2010]. Other relevant social emotions such as reproach, regret, anger, and disappointment, also appear to have similarly non-negligible effect on social behavior, especially in the context of those sequential games.

While we have demonstrated that the choice of a particular type of social interaction largely depends on the scientific objective (for an exhaustive list of well known games used in experimental economics, see Camerer [2003]; Roth and Kagel [1995]), one should note that, as for any other experiment from the physical sciences, the methodology for designing an experiment is crucial to the success of the corresponding economic study. In fact, as game theory considers social interactions through abstract representations, an important step is to specify a set of sufficiently explicit rules or instructions so that the participants actually play the game as they are intended to. More precisely, one may argue that the two main factors that need to be controlled through experiments are the individuals’ preferences and uncertainty.

Concerning preferences, the subjects involved in economic experiments are generally incentivized with real money payoffs. Such financial incentives, which are rarely used in other scientific areas (e.g. psychology), have the advantage to be, in average, similarly valued by human agents. Moreover, past experiments clearly indicate that, in general, subjects participate more seriously and show more effort with such incentives. Using monetary payoffs therefore represents the best known method for manipulating people’s preferences over outcomes.

Regarding the other issue of controlling the subjects’ uncertainty, things are not so simple. First, the instructions of a game must be determined through a careful choice of wording that lies at the right level of abstraction: a too abstract description of the game can clearly obscure the situation to the players, whereas too specific instructions can bias their perception of the game. Similarly, one also has to control the subjects’ available information about what each other knows. For example, such common knowledge of the game itself can be reached through reading the game’s instructions out loud in front of all the participants. Furthermore, an individual’s knowledge about the identity of the other participants is also a particularly important issue. In most experiments, anonymity is assured between subjects, as a means to avoid any reputation effect. Similarly, in the case of several interactions, the subjects must also be informed about any possible change in the identity of the partner(s): e.g., in order to avoid any learning effect,
the participants should know that they cannot be matched with the same partner more than once. However, one should clearly admit that control over one’s actual beliefs is a difficult task. This is why many experiments often make use of complementary questionnaires, as a means to provide some additional information about the subjects’s background, their understanding of the game, their beliefs, and occasionally their feelings and emotions.

Given the complexity of designing such economic experiments, it is worth noting that this methodology has been subject to various criticisms in the past, which denounce a lack of realism in making decisions in a lab. More specifically, those objections include the “strangeness” of the game designs representing situations that are rarely met in real life. For example, while many experiments involve one-shot anonymous interactions, one may claim that most daily life situations are concerned with repeated interactions between people that often know each other. Although such an argument can hardly be denied (we will come back to it in the next section), one should note that one-shot anonymous games are however not rare outcomes, especially within advanced market societies: as argued in Gintis [2009], most interactions we have with strangers in daily life are of this form. Moreover, as pointed out in Camerer [2011], it must be clarified that the primary concern of economic experimentation is not to generalize from the lab to the field, it is instead “to establish a general theory linking economic factors, such as incentives, rules, and norms, to behavior”.

2.4.2 Other-regarding preferences

So far in this chapter, we have assumed very little about what defines an agent’s preferences. In fact, in Section 2.1, we have only described the constraints underlying rational preferences, and presented a way to measure those preferences (through either ordinal or cardinal utility functions). However, those theories say nothing about the different types of preferences one may hold. We therefore attempt to fill this gap throughout this section.

The most common assumption that has been made in many existing economic theories is to consider human beings as self-centered agents that aim at reaching the best possible outcome for satisfying their own primary desires. Those desires then include all sorts of physical or psychological needs that can improve one’s own welfare. This type of material preferences often refers to the well known principle of homo economicus in the literature. As shown through Section 2.2, in the context of social interactions, such a self-regarding rational agent thus cares about other players’ preferences and behavior only insofar as these impact his own payoff. Formally, given a set of agents $Agt$ and a set of outcomes $O$, let us define a material payoff function $\pi_i$ that specifies, for each agent $i \in Agt$ and each outcome $o \in O$, a measure of how much $o$ is good or bad to $i$. 

42
\[ \pi_i : O \rightarrow \mathbb{R} \]

Note that this function simply corresponds to the measure of agent \( i \)'s cardinal utility, as specified in Section 2.1.2.

However, while it is clear that this type of individualistic behavior can actually be followed in some particular situations (e.g., in various types of markets), it also appears to be largely ignored in many everyday interactions. As an example, suppose that I am queuing at a cinema ticket center. I may then want to cut the line to buy my ticket because waiting is materially more costly to me than not waiting (e.g., it makes me bored). Such a behavior then appears to be the only rational move for me: given everybody else stays in line, it is in my best interest to cut the line. Yet, I may eventually choose not to do it because it would not be fair to other people in the queue. Is this behavior then irrational? Our claim is that irrationality is not the issue in this case. Instead, my preferences may simply be sensitive to the welfare of other individuals involved in the situation. In other words, my utility does not only rely on my own material payoffs, it also depends on others’ material payoffs. Formally, for every agent \( i \in \text{Agt} \), \( i \)'s utility function \( u_i \) can be defined such that, for every outcome \( o \in O \), the value \( u_i(o) \) corresponds to an aggregation of the material payoffs of every agent from \( \text{Agt} \) (i.e., \( \pi_j(o) \), for every \( j \in \text{Agt} \)).

This type of utility function is said to specify one’s other-regarding preferences (also called social preferences). However, one may wonder about the possible definitions of the above utility function \( u_i \). Let us therefore briefly discuss various well known theories that have already been proposed in the economics literature. In Fehr and Schmidt [1999], Fehr & Schmidt have provided empirical evidence that people are genuinely sensible to equality, that is, an individual dislikes being either better off or worse off than any other person. As a result, they provide a concrete expression of the above function \( u_i \) that satisfies such a theory of inequity aversion. More details about this theory can be found in Section 4.5.1 from Chapter 4. Similarly, in Charness and Rabin [2002], Charness & Rabin present an alternative model of fairness, which relies on, what they call, quasi-maximin preferences. More specifically, they assert that one’s utility is defined by a combination of one’s own material payoffs, and the social welfare of the group as a whole. According to their theory, such a social welfare is determined as a trade-off between satisfying the worst off individual and maximizing the total utility of the group (such preferences are inspired by Rawls’ theory of justice from Rawls [1971]). More details regarding this alternative interpretation of function \( u_i \) can also be found in Section 4.5.2 from Chapter 4. Furthermore, in Levine [1998], Levine introduces another theory of social preferences that is based upon the concept of altruism. More precisely, he defines an individual’s adjusted utility according to one’s propensity to be
altruistic/selfish/spiteful, and one’s regard for altruistic/selfish/spiteful opponents.

Through the previous theories of social (or other-regarding) preferences, which are all supported by experimental data, one can observe that they do not contradict the classical theory of rationality that we have been exploring so far. In fact, once an individual’s utility is correctly adjusted according to either of the above interpretations, then the rationality principle can still hold. In other words, those models explain deviation from self-regarding behavior, not by rejecting rationality, but instead by refining one’s preferences, which may simply be influenced by the welfare of other individuals involved in the interaction. Those results therefore lead to the relevant distinction that can be made between being self-regarding and being self-interested: as pointed out in Gintis [2009], while it makes sense to say that a self-regarding agent is necessarily self-interested, the reverse may not always hold. For example, I may indeed get great pleasure from offering you a gift that makes you particularly happy, even though it is costly for me to do so. In this particular case, although I am clearly not self-regarding, I am however self-interested in the sense that the increase of your welfare somehow also contributes to improving my own well-being.

According to the previously mentioned studies, it is worth summarizing the general result as follows: any given individual tends to rank outcomes according to everybody’s social preferences in the actually occurring outcome. Formally, if some other-regarding agent $i \in Agt$ is asked to subjectively compare two outcomes $o_1 \in O$ and $o_2 \in O$, then $i$ will consider the material payoff functions of every other agent from $Agt$ when applied to these two outcomes. More precisely, instead of simply comparing $\pi_i(o_1)$ with $\pi_i(o_2)$ as a self-regarding agent, $i$ will instead compare $u_i(o_1)$ with $u_i(o_2)$. However, this points out to an important remark: in comparing the two outcomes, agent $i$ may also actually take into account preferences related to other alternative outcomes that could have occurred. In order to illustrate this point, let us consider the following game that involves two players (Alice and Bob). As for the Hi-Lo matching game presented earlier (see Figure 2.7), both players have the same choice set, that is, each can choose between some option $A$ and some option $B$. The game is also symmetric, in the sense that no matter the actual outcome, both agents will necessarily earn the same material payoff. The corresponding payoff matrix (i.e., $\pi_{Alice}$ and $\pi_{Bob}$) is depicted in Figure 2.9.

According to this game, if both players select $A$, then they both get 10 €. If they both play $B$, then they each get 5 €. However, if Alice plays $A$ and Bob plays $B$, then they each get 6 €. Similarly, whenever Alice plays $B$ while Bob plays $A$, then each gets 50 €.

In this context, let us now look at Alice’s preferences in more details through the following question. Does Alice prefer outcome $(A, A)$ over outcome $(B, B)$? A positive answer looks straightforward as $\pi_{Alice}(A, A) > \pi_{Alice}(B, B)$. Moreover, note
that caring about Bob’s desires according to any of the above theories of social preferences cannot change Alice’s preferences. In other words, for any relevant function $u_{Alice}$ that defines Alice’s utility, we similarly have that $u_{Alice}(A, A) > u_{Alice}(B, B)$. However, one should note that there exists another intuitive interpretation to the way Alice views this problem. If outcome $(A, A)$ actually occurs, Alice would realize that she could have allowed a much higher payoff (for the group, Bob and herself) had she played $B$ instead of $A$. This negative feeling may then considerably lower Alice’s utility for this outcome. On the other hand, if outcome $(B, B)$ actually occurs, then Alice does not have such a strong negative feeling as she could not have reached a much better payoff for anybody (or for the group) by behaving differently. As a result of this interpretation, Alice’s utility may actually be lower for outcome $(A, A)$ than for outcome $(B, B)$.

As a means to justify this alternative interpretation, one may consider a well-known social emotion: regret. Indeed, psychologists and economists (e.g., Loomes and Sugden [1982]; Sugden [1985]; Zeelenberg [1999]; Zeelenberg et al. [1998]) agree that regret is a negative, cognitively determined emotion that we experience when realizing or imagining that our present situation would have been better, had we acted differently. Applied to the above scenario, Alice can feel such regret in the case of outcome $(A, A)$. On the other hand, the trigger of such a social emotion is reasonably expected to be negligible in the case of outcome $(B, B)$ (i.e., payoffs are not very different in outcome $(B, B)$ and in outcome $(A, B)$). Consequently, if Alice is sufficiently regret averse, she will then prefer outcome $(B, B)$ over outcome $(A, A)$.

More generally, the above game from Figure 2.9 illustrates the fact that one’s utility for a given outcome $o$ may not only rely on everybody’s material payoffs related to $o$, but also on those material payoffs related to other alternative outcomes from $O$.

Formally, given some agent $i \in Agt$, one can refine $i$’s utility function $U_i$ such that, for every outcome $o \in O$, the value $U_i(o)$ depends on the material payoffs.

![Figure 2.9: Game with possible regret](image-url)
of every agent from \( Agt \) in every possible outcome from \( O \) (i.e., \( \pi_j(o') \), for every \( j \in Agt \) and every \( o' \in O \)).

Furthermore, it is worth indicating that all the above theories of other-regarding preferences rely on some properties that are intrinsic to the individuals. In fact, they claim that different individuals may naturally have different levels of fairness, altruism, inequity aversion, or regret aversion. However, those theories remain vague about whether an individual’s sensitivity to such attitudes depends on the individuals’ identities. In other words, are the identities of the persons one interacts with irrelevant to defining one’s social preferences? The above models can clearly not answer this question as they strictly apply to situations of anonymous interactions. Nevertheless, as previously mentioned in Section 2.4.1, it is clear that many real-life situations involve a high number of interactions between individuals that are socially tied with one another. Intuitively, it therefore makes sense to state that one often behaves differently in the same situation depending on whether the persons involved are good friends or perfect strangers.

This leads us to claim that knowing the identities of the persons one interacts with also directly affect one’s own preferences. In other words, when making a decision in a social situation, one may value other individuals’ interests differently depending on their level of social closeness. The main consequence of this theory of social ties is that it still does not conflict with the classical concept of rationality, it simply refines it further. In fact, a rational behavior still consists in maximizing one’s own preferences, given one’s beliefs about others’ choices. The only difference with classical economic theory lies in those preferences that are not only composed of one’s own material interest, but also incorporate one’s concern for the relative welfare of other socially tied individuals. As a result, we introduce the concept of social rationality to denote the type of rationality displayed by individuals in multi-agent decision-making situations that involve various types of social relationships.

In order to formally characterize this notion of social rationality, we will first present in Chapter 4 the study of a particular type of two-player coordination game with the following interesting properties: the interactive situation involves a problem of co-operation similar to that discussed in Section 2.4.1 (see Figure 2.7). Furthermore, it allows the participants to face a particular social dilemma, which somewhat also resembles that from the trust game presented in Figure 2.8. Following such a theoretical study, Chapter 5 introduces the corresponding empirical analysis, which provides evidence of the actual effects of social ties on human behavior, and justify the need for defining an alternative theoretical model. The resulting generalized model will then be presented and extensively studied through Chapter 8.
Chapter 3

Epistemic Rationality in Dynamic Games

“The doorstep to the temple of wisdom is a knowledge of our own ignorance.”
— Benjamin Franklin

“It is only when we forget all our learning that we begin to know.”
— Henry David Thoreau
Journal, 4 October 1859

The aim of this chapter is to propose a modal logic framework that allows to reason about epistemic games in extensive form. In this kind of games, players decide what to do according to some general principles of rationality while being uncertain about several aspects of the interaction such as other agents’ choices, other agents’ preferences, etc. We therefore propose both a semantic and a syntactic analysis of such sequential games in modal logic. In particular, we introduce a multi-modal logic interpreted on a Kripke-style semantics which integrates the concepts of action, strategy, knowledge and preference and which allows to reason about the properties of extensive form games. In order to illustrate the expressive power of the logic, we define in its object language the well known concepts of rationality and backward induction, as they are defined according to economic theory. Based on these definitions, we then provide a syntactic proof of Aumann’s theorem that states the following: “for any non degenerate game of perfect information, common knowledge of rationality implies the backward induction solution” (Aumann [1995]). While there exist other logics that formalize similar theorems, none of these is expressive enough to provide syntactic proofs that would emphasize the
various requirements assumed for the theorems. For example, while Baltag et al. [2009] presents a logic that can correctly define the statement of Aumann’s theorem, no syntactic proof of it is provided, and its language does not allow to verify whether the theorem holds when the epistemic conditions are weakened. Indeed, if one realistically only considers common knowledge to be bounded to some finite level, then the maximal depth of the game represents an important variable to the proof of the theorem. By considering the temporal dimension of such extensive form games, we demonstrate its relevance to the proof of some weaker version of the theorem.

More generally, our intention, throughout this chapter, is not to show that a syntactic derivation of Aumann’s theorem is interesting in itself. Instead, we wish to demonstrate that this kind of analysis is useful to identify specific assumptions about the relationship between players’ knowledge and the game structure that are needed in order to prove the theorem.

The syntactic proof of Aumann’s theorem is given in Appendix A.

### 3.1 A modal logic of actions, strategies, knowledge and preferences

We present in this section the modal logic ELEG (Epistemic Logic of Extensive Games) integrating the concepts of action, strategy, knowledge and preference. This logic supports reasoning about games in extensive form in which an agent might be uncertain about the other agents’ current and future choices of actions.

#### 3.1.1 Syntax

The syntactic primitives of the logic ELEG are the finite set of agents \( \text{Agt} \), the set of atomic propositions \( \text{Atm} \), a nonempty finite set of atomic action names \( \text{Act} = \{\alpha_1, \alpha_2, \ldots, \alpha_{|\text{Act}|}\} \), a non-empty finite set of \( N \) integers \( I = \{0, \ldots, N\} \).

The language \( \mathcal{L} \) of the logic ELEG is given by the following BNF (Backus-Naur Form) grammar:

\[
\begin{align*}
\chi & ::= \ p \mid \alpha \mid \text{turn}_i \mid \text{end} \mid k_i \\
\phi & ::= \chi \mid \neg \phi \mid \phi \lor \phi \mid \square \phi \mid \text{AX}\phi \mid [k_i]\phi \mid \text{X}\phi
\end{align*}
\]

where \( p \) ranges over \( \text{Atm} \), \( i \) ranges over \( \text{Agt} \), \( \alpha \) ranges over \( \text{Act} \), and \( k \) ranges over \( I \). Formulas \( \chi \) are called atomic formulas. The classical Boolean connectives \( \bot, \top, \land, \rightarrow \) and \( \leftrightarrow \) are defined from \( \lor \) and \( \neg \) in the usual manner.

The formula \( \alpha \) has to be read “the action \( \alpha \) is performed”, while \( \text{turn}_i \) and \( k_i \) are read respectively “it is agent \( i \)’s turn to play”, and “the current strategy profile
will ensure a payoff $k$ to agent $i$. Finally, $\text{end}$ is meant to stand for “the current vertex of the game is an end vertex”.

The operator $\Box$ is used to quantify over strategy profiles of the current game. $\Box \varphi$ has to be read “$\varphi$ holds for all strategy profiles of the current extensive game”. The operator $AX$ is used to quantify over next vertices of the current extensive game. $AX \varphi$ has to be read “$\varphi$ is true at every possible next vertex along the current strategy profile”.

The formula $[K_i] \varphi$ is read as usual “agent $i$ knows that $\varphi$ is true”. $\langle \rangle \varphi$ is the standard temporal operator of next. The formula $X \varphi$ has to be read “$\varphi$ will be true next”.

Moreover, the following abbreviations are given:

$\Diamond \varphi \overset{\text{def}}{=} \lnot \Box \lnot \varphi$

$EX \varphi \overset{\text{def}}{=} \lnot AX \lnot \varphi$

$\langle K_i \rangle \varphi \overset{\text{def}}{=} \lnot [K_i] \lnot \varphi$

$\alpha_i \overset{\text{def}}{=} \alpha \land \text{turn}_i$

$X^0 \varphi \overset{\text{def}}{=} \varphi$

$X^{n+1} \varphi \overset{\text{def}}{=} XX^n \varphi$

$AX^0 \varphi \overset{\text{def}}{=} \varphi$

$AX^{n+1} \varphi \overset{\text{def}}{=} AX(AX^n \varphi)$

$AX^{\leq n} \varphi \overset{\text{def}}{=} \bigwedge_{0 \leq m \leq n} AX^m \varphi$

$EX^n \varphi \overset{\text{def}}{=} \lnot AX^n \lnot \varphi$

$EX^{\leq n} \varphi \overset{\text{def}}{=} \lnot AX^{\leq n} \lnot \varphi$

$\Diamond \varphi$ has to be read “$\varphi$ holds for at least one strategy profile of the current extensive game”, whereas $EX \varphi$ has to read “$\varphi$ is true in at least one possible next vertex along the current strategy profile”. $\langle K_i \rangle \varphi$ has to be read “agent $i$ thinks that $\varphi$ is possible”, whereas $\alpha_i$ has to be read “agent $i$ performs the action $\alpha$”. $X^n$ has to be read “$\varphi$ will be true $n$ steps from now”. Operators $AX^n \varphi$ and $AX^{\leq n} \varphi$ respectively read “$\varphi$ is true in every vertex that can be reached in exactly $n$ step(s) from now, along the current strategy profile” and “$\varphi$ is true in every vertex that can be reached within $n$ step(s) from now, along the current strategy profile”. Finally the corresponding dual operators $EX^n \varphi$ and $EX^{\leq n} \varphi$ can be interpreted as “$\varphi$ is true in at least one vertex that can be reached in exactly $n$ step(s) from now, along the current strategy profile” and “$\varphi$ is true in at least one vertex that can be reached within $n$ step(s) from now, along the current strategy profile”.

As common in Propositional Dynamic Logic (PDL), we introduce an operator of sequential composition “;”. We define the set $\text{Seq}$ of action sequences as the smallest set such that: $\alpha \in \text{Seq}$ for any $\alpha \in \text{Act}$, and if $\epsilon_1, \epsilon_2 \in \text{Seq}$ then $\epsilon_1; \epsilon_2 \in \text{Seq}$. Moreover, we consider $\text{Seq}^n \subseteq \text{Seq}$ to be the set of action sequences of length $n$. The fact that a given action sequence will occur and $\varphi$ will be true afterwards can be defined in the object language by means of the following definition:

$\langle \alpha_0; \ldots ; \alpha_n \rangle \varphi \overset{\text{def}}{=} \bigwedge_{0 \leq i \leq n} X^i \alpha_i \land X^n \varphi$
We use \([\text{EK}_C]\varphi\) as an abbreviation of \(\Lambda_{i \in C}[K_i]\varphi\), i.e., every agent in \(C\) knows \(\varphi\) (if \(C = \emptyset\) then \([\text{EK}_C]\varphi\) is equivalent to \(\top\)). Then we define by induction \([\text{EK}_C^k]\varphi\) for every natural number \(k \in \mathbb{N}\):

\[
[\text{EK}_C^0]\varphi \stackrel{\text{def}}{=} \varphi
\]

and for all \(k \geq 1,

\[
[\text{EK}_C^k]\varphi \stackrel{\text{def}}{=} [\text{EK}_C][([\text{EK}_C^{k-1}]\varphi)]
\]

Similarly, we define for all natural numbers \(n \in \mathbb{N}\):

\[
[\text{CK}_C^n]\varphi \stackrel{\text{def}}{=} \varphi
\]

and for all \(n \geq 1,

\[
[\text{CK}_C^n]\varphi \stackrel{\text{def}}{=} \bigwedge_{1 \leq k \leq n} [\text{EK}_C^k]\varphi
\]

\([\text{CK}_C^n]\varphi\) expresses \(C\)'s common knowledge that \(\varphi\) up to \(n\) iterations, i.e., everyone in \(C\) knows \(\varphi\), everyone in \(C\) knows that everyone in \(C\) knows \(\varphi\), and so on until level \(n\).

3.1.2 Semantics

A strategic structure includes a set of vertices, a set of strategy profiles, a successor function associating vertices and strategy profiles to vertices, a turn-taking function assigning agents to vertices. The set of vertices includes end vertices.

**Definition 3.1 (Strategic structure)** A strategic structure is a tuple \(T = (V, \mathcal{Q}, S, \text{next}, \text{EndV})\) where:

- \(V\) is a non-empty set of vertices;
- \(\mathcal{Q}\) is a total function \(\mathcal{Q} : V \rightarrow \text{Agt}\) mapping vertices to agents;
- \(S\) is a non-empty set of strategy profiles on \(V\), and every strategy profile \(s \in S\) is a total function \(s : V \rightarrow \text{Act}\) mapping vertices to actions;
- \(\text{next}\) is a partial function \(\text{next} : V \times S \rightarrow V\) mapping vertices and strategy profiles to vertices such that:
  
  \(C1\) if \(s(w) = s'(w)\) then \(\text{next}(w, s) = \text{next}(w, s')\);
- \(\text{EndV} \subseteq V\) is the set of end vertices such that:
  
  \(C2\) \(w \in \text{EndV}\) if and only if, \(\text{next}(w, s)\) is undefined for every \(s\).
\( Q(w) = i \) means that at vertex \( w \) it is agent \( i \)'s turn to play, and \( \text{next}(w, s) = w' \) means that \( w' \) is the next vertex of \( w \) with respect to the strategy profile \( s \). We call \textit{index} a pair \((w, s)\) with \( w \in V \) and \( s \in S \). We define \( H = V \times S \) the set of all indices.

Consider every vertex of the game and is not restricted to a single player’s moves as usually done in game theory. Moreover, for every \( s \in S \), a single player \( i \)'s strategy \( s_i \) can be defined as the restriction of \( s \) to the vertices in which it is agent \( i \)'s turn to play.

According to the Constraint C1, two strategy profiles selecting the same action at a given vertex lead to the same next vertex. According to the constraint C2, an end vertex is a vertex which does not have a next vertex.

**Definition 3.2 (Successor)** \( R \) is a relation on vertices such that:

\[ \text{for every } w, v \in V, \ wRv \text{ if and only if there is } s \in S \text{ such that } \text{next}(w, s) = v. \]

\( wRv \) means that vertex \( v \) is a successor of vertex \( w \).

An extensive game model is nothing but a strategic structure supplemented with accessibility relations for agents’ knowledge over strategy profiles, agents’ preferences and a valuation of atomic propositions.

**Definition 3.3 (Extensive game model)** An extensive game model is a tuple \( M = \langle T, \{ E_i \mid i \in \text{Agt} \}, \{ P_i \mid i \in \text{Agt} \}, \pi \rangle \) where:

- \( T \) is a strategic structure;
- every \( E_i \) is an equivalence relation on \( S \) such that:
  - \( C3 \) if \( sE_is' \) and \( Q(w) = i \), then \( s(w) = s'(w) \);
- every \( P_i \) is a total function \( P_i : H \rightarrow \mathbb{I} \) mapping every index to an integer such that:
  - \( C4 \) if \( \text{next}(w, s) = w' \), then \( P_i(w, s) = k \) if and only if \( P_i(w', s) = k \);
  - \( C5 \) if \( w \in \text{End}V \) and \( s(w) = s'(w) \) then \( P_i(w, s) = P_i(w, s') \).
- \( \pi : \text{Atm} \rightarrow 2^H \) is a valuation function on indices.

\( sE_is' \) means that agent \( i \) cannot distinguish strategy profile \( s \) from the strategy profile \( s' \). \( P_i(w, s) = k \) means that the strategy profile \( s \) played at the vertex \( w \) will ensure a payoff \( k \) to agent \( i \).

Constraint C3 is the assumption that every agent knows his choice when it is his turn to play Aumann [1995]; Battigalli and Bonanno [1999]. Constraint C4 correctly expresses the fact that in an extensive form game, preferences are built over
histories, where a history is nothing but a sequence of indices \((w_0, s), \ldots, (w_n, s), \ldots\) such that \(\text{next}(w_i, s) = w_{i+1}\) for every \(0 \leq i\). According to Constraint C5, two strategies selecting the same action at an end vertex lead to the same payoff for an agent. In other words, at an end vertex the payoff of an action is uniquely determined.

**Example** In order to illustrate the use of our logic ELEG to model extensive form games, let us consider a well known game in economics, namely the trust game (Berg et al. [1995]). The binary version of the trust game (BTG) involves two players, the truster (Alice) and the trustee (Bob), playing sequentially in the following way: first Alice can choose between leaving the game and divide the amount of 2€ equally with Bob (i.e., 1€ for each) or let Bob play. In the latter case, Bob can either divide the amount of 6€ equally with Alice (i.e., 3€ for each) or keep the whole amount for himself (i.e., 6€ for himself and 0€ for the truster). Consider a version of this game, whose graphical representation is depicted in Figure 3.1.

---

![Binary Trust Game (BTG)](image)

**Figure 3.1: Binary Trust Game (BTG)**

In Figure 3.1, let us consider *Alice*, as the truster who plays at vertex \(v\), and *Bob*, as the trustee who plays at vertex \(w\). At each leaf of the tree, payoffs take the form \((X, Y)\), where *Alice* gets \(X\)e and *Bob* gets \(Y\)e. Moreover, actions named \(C_a\) and \(D_a\) respectively stand for “*Alice* cooperates” and “*Alice* defects”. Similarly, actions named \(C_b\) and \(D_b\) respectively stand for “*Bob* cooperates” and “*Bob* defects”.

Therefore, we suppose \(\text{Agt} = \{\text{Alice, Bob}\}\), \(\text{Act} = \{C_a, C_b, D_a, D_b\}\), \(V = \{v, w\}\), \(\text{EndV} = \{w\}\), and \(S = \{s_1, s_2, s_3, s_4\}\);

Let us now represent the extensive game model corresponding to the binary trust game in ELEG:

- \(\text{next}(v, s_1) = w, \text{next}(v, s_2) = w;\)
• \( Q(v) = Alice, Q(w) = Bob; \)

• \( E_{Alice}(s_1) = E_{Bob}(s_2) = \{s_1, s_2\}; E_{Alice}(s_4) = E_{Alice}(s_3) = \{s_3, s_4\}; \)

• \( E_{Bob}(s_1) = E_{Bob}(s_3) = \{s_1, s_3\}; E_{Bob}(s_2) = E_{Bob}(s_4) = \{s_2, s_4\}; \)

• \( P_{Alice}(v, s_1) = P_{Alice}(w, s_1) = P_{Alice}(w, s_3) = 3, \)
  \( P_{Bob}(v, s_1) = P_{Bob}(w, s_1) = P_{Bob}(w, s_3) = 3, \)
  \( P_{Alice}(v, s_2) = P_{Alice}(w, s_2) = P_{Alice}(w, s_4) = 0, \)
  \( P_{Alice}(v, s_3) = P_{Alice}(v, s_4) = P_{Bob}(v, s_3) = P_{Bob}(v, s_4) = 1, \)
  \( P_{Bob}(v, s_2) = P_{Bob}(w, s_2) = P_{Bob}(w, s_4) = 6 \)

This model represents the four possible strategy profiles \( s_1, s_2, s_3 \) and \( s_4 \) of the BTG, each of which includes the same two vertices \( v \) and \( w \) where various actions occur:

• \( s_1 \) corresponds to strategy profile \((C_a, C_b)\);

• \( s_2 \) corresponds to strategy profile \((C_a, D_b)\);

• \( s_3 \) corresponds to strategy profile \((D_a, C_b)\);

• \( s_4 \) corresponds to strategy profile \((D_a, D_b)\);

Vertices \( v \) and \( w \) represent the nodes within the game where respectively Alice and Bob have to play. The epistemic relations \( E_{Alice} \) and \( E_{Bob} \), as they are defined in this model, represent perfect uncertainty for each player over the strategy profiles. One should note however that these epistemic relations are only examples and could possibly be defined differently without modifying the strategic structure of the game.

**Definition 3.4 (Truth conditions)** Truth of a formula in a model \( M \) at a given index \((w, s)\) is defined as follows:

• \( M, w, s \models p \iff (w, s) \in \pi(p); \)

• \( M, w, s \models \neg \varphi \iff M, w, s \not\models \varphi; \)

• \( M, w, s \models \varphi \lor \psi \iff M, w, s \models \varphi \) or \( M, w, s \models \psi; \)

• \( M, w, s \models \alpha \iff s(w) = \alpha; \)

• \( M, w, s \models \text{turn}_i \iff Q(w) = i; \)

• \( M, w, s \models \text{end} \iff w \in \text{EndV}; \)

• \( M, w, s \models k_i \iff P_i(w, s) = k; \)
\[ M, w, s \models X \varphi \iff \text{next}(w, s) \text{ is defined then } M, \text{next}(w, s), s \models \varphi; \]
\[ M, w, s \models \Box \varphi \iff M, w, s' \models \varphi \text{ for all } s' \in S; \]
\[ M, w, s \models AX \varphi \iff M, w', s \models \varphi \text{ for all } w' \in V \text{ such that } w \mathcal{R} w'; \]
\[ M, w, s \models [K_i] \varphi \iff M, w, s' \models \varphi \text{ for all } s' \text{ such that } s \mathcal{E}_i s'. \]

A formula \( \varphi \) is **true in an extensive game model** \( M \) iff \( M, w, s \models \varphi \) for every vertex \( w \) in \( V \) and every strategy profile \( s \) in \( S \). \( \varphi \) is **ELEG-valid** (noted \( \models \varphi \)) iff \( \varphi \) is true in all extensive game models. \( \varphi \) is **ELEG-satisfiable** iff \( \neg \varphi \) is not ELEG-valid.

### 3.1.3 Some validities

Table 3.1 provides an exhaustive list of ELEG validities that will be sufficient to provide in Section 3.3 a syntactic proof of Aumann’s theorem.

Let us prove the validity \( \text{Perm}_{[K_i]AX} \) as an example. Assume \( M, w, s \models [K_i]AX \varphi \) for an arbitrary ELEG model \( M \). This is equivalent to say that \( M, w, s' \models AX \varphi \) for all \( s' \) such that \( s \mathcal{E}_i s' \) which, in turn, is equivalent to say that \( M, w', s' \models \varphi \) for all \( (w', s') \) such that \( s \mathcal{E}_i s' \) and \( w \mathcal{R} w' \). The latter is equivalent to say that \( M, w', s \models [K_i] \varphi \) for all \( w' \) such that \( w \mathcal{R} w' \) which, in turn, is equivalent to say that \( M, w, s \models AX[K_i] \varphi \).

In the sequel, we will write \( \Vdash_{\text{ELEG}} \varphi \) to mean that \( \varphi \) can be derived by means of the list of principles given in Table 3.1. The study of a complete axiomatization of the logic ELEG is postponed to future work.

### 3.2 Backward induction and rationality

We here define two fundamental concepts in Aumann’s epistemic analysis of extensive form games: the concept of backward induction and the concept of rationality.

As a matter of simplicity to later prove Aumann’s Theorem, we only provide in this section simplified formal definitions that only apply to games of uniform depth. One should note however that more general definitions of both backward induction and rationality can easily be expressed in ELEG.

In order to define the concept of backward induction, we first introduce the well known Nash equilibrium, which can also be expressed in ELEG.

#### 3.2.1 Nash equilibrium

The Nash equilibrium is a well known solution concept, according to which every player chooses a strategy such that none of them can benefit by individually
Table 3.1: Some validities of ELEG
changing his own strategy. In other words, every agent plays a strategy that is a
best response to others. In order to define this concept for extensive form games,
one first needs to express the following notion of Best Response (BR) in ELEG.

For the case \( n = 0 \) we define (for \( i \in \text{Agt}, k \in I \)):

\[
\text{BR}^0_i (k) \equiv \text{end} \land (\text{turn}_i \rightarrow \square \bigvee_{h \in I : h \leq k} h_i)
\]

For every \( n > 0 \) we define (for \( i \in \text{Agt}, k \in I \)):

\[
\text{BR}^n_i (k) \equiv \text{XBR}^{n-1}_i (k) \land (\text{turn}_i \rightarrow \text{AX} \bigvee_{h \in I : h \leq k} h_i \land \text{BR}^{n-1}_i (k))
\]

This notion of best response therefore reads as follows: \( M, w, s \models \text{BR}^n_i (k) \) if
and only if agent \( i \) cannot obtain a payoff higher than \( k \) by individually deviating
from his current strategy from \( s \), when starting from vertex \( w \) from the finite game
of uniform depth \( n \) (assuming other agents play their own strategies from \( s \)).

Based on the previous concept, the corresponding formal definition in ELEG
of the Nash solution in a finite game of uniform depth \( n \) is as follows.

\[
\text{Nash}^n \equiv \bigwedge_{i \in \text{Agt}} \bigvee_{k \in I} k_i \land \text{BR}^n_i (k)
\]

In this case: \( M, w, s \models \text{Nash}^n \) if and only if the current strategy profile \( s \),
when starting from vertex \( w \), corresponds to a Nash solution for the finite game
of uniform depth \( n \).

One should note that this concept of Nash equilibrium implies that each chosen
action at every vertex along the actual path does maximize the player’s payoff.

**Theorem 3.1** For every \( n \in \mathbb{N} \), we have:

\[
\vdash_{\text{ELEG}} \text{Nash}^{n+1} \rightarrow \text{XNash}^n
\]

However, this definition does not guarantee local best response at vertices that
are not on this path.

\[
\not\vdash_{\text{ELEG}} \text{Nash}^{n+1} \rightarrow \text{AXNash}^n
\]

In other words, this definition implies that a best response action may rely
on some alternative paths that are not necessarily local best responses themselves.
Such a weakness of the Nash equilibrium in the context of extensive games is known
in the economic literature through the concept of non-credible threats: the Nash
solution can indeed consider some threats that rational players would actually not
carry out, because it would not be in their best interest to do so. In order to
rule out such threats, one then needs to consider the Nash equilibrium not only at
every vertex along the actual path, but also at every vertex along every alternative
path of the game. Such a strengthening corresponds to the concept of backward
induction solution that is defined below.
3.2.2 Backward induction

The notion of backward induction represents the process of reasoning backwards in time, starting from each end vertex of the game in order to determine a sequence of optimal actions. This method is generally used to compute the subgame perfect Nash equilibria in sequential games. The backward induction (BI) solution in a game of depth \( n \) (i.e., where at most \( n \) steps are necessary to reach an end vertex of the game) can be computed by iterating the process \( n \) times, as the BI solution at one state relies on the BI solution at every possible successive state. Therefore, the first step BI solution \((n = 0)\) corresponds to the maximization of preferences for the last player to play at each possible end vertex of the game. The BI solution after \( n + 1 \) \((n > 0)\) steps corresponds to the maximization of the current player’s preferences, considering only those that satisfy the BI solution after \( n \) steps at any possible next state.

The recursive formal definition in ELEG of the BI solution in a finite game of uniform depth \( n \) is as follows.

For the case \( n = 0 \) we define:

\[
\text{BI}^0 \overset{\text{def}}{=} \text{end} \land \bigvee_{i \in \text{Agt}, k \in I} \text{turn}_i \land k_i \land \Box \left( \bigvee_{h \in I : h \leq k} h_i \right)
\]

For every \( n > 0 \) we define:

\[
\text{BI}^n \overset{\text{def}}{=} \neg \text{end} \land \bigvee_{i \in \text{Agt}, k \in I} \text{turn}_i \land k_i \land \text{AX}(\text{BI}^{n-1} \land \bigvee_{h \in I : h \leq k} h_i)
\]

Therefore: \( M, w, s \models \text{BI}^n \) if and only if the current strategy profile \( s \), when starting from vertex \( w \), corresponds to a backward induction solution in the sub-game of uniform depth \( n \), which can be computed in \( n + 1 \) steps.

Note that without any additional constraint on the game involved, none of these definitions does guarantee that a unique BI solution will be found after each step.

Moreover, as the backward induction computes the subgame perfect Nash equilibrium, the following formula therefore becomes valid in ELEG:

**Theorem 3.2** For every \( n \in \mathbb{N} \), we have:

\[
\vdash_{\text{ELEG}} \text{BI}^n \leftrightarrow \bigwedge_{0 \leq m \leq n} \text{AX}^m \text{Nash}^{n-m}
\]

One should also note that the backward induction simply corresponds to a restriction on the Nash equilibrium:

\[
\vdash_{\text{ELEG}} \text{BI}^n \rightarrow \text{Nash}^n
\]
3.2.3 Epistemic rationality

The following ELEG definition characterizes a notion of material rationality that is supposed in Aumann’s epistemic analysis of extensive form games:

\[ \text{Rat}^\text{end}_i \overset{\text{def}}{=} (\text{end} \land \text{turn}_i) \rightarrow \bigvee_{k \in I} (k_i \land \Box(\bigvee_{h \in I, h \leq k} h_i)) \]

\(\text{Rat}^\text{end}_i\) means that an agent \(i\) is rational at an end vertex (i.e., at some end vertex of the game) if and only if \(i\) chooses an action that maximizes his individual payoff. Note that in this case, rationality does not rely on any epistemic component.

\[ \text{Rat}^{\neg\text{end}}_i \overset{\text{def}}{=} (\neg\text{end} \land \text{turn}_i) \rightarrow \bigvee_{k \in I} (K_i(k_i \land \AX(\bigvee_{h \in I, h \leq k} (K_i h_i)))) \]

\(\text{Rat}^{\neg\text{end}}_i\) means that an agent \(i\) is rational at any intermediate vertex (any node that is not an end vertex of the game) if and only if \(i\) chooses an action in such a way that what he considers possible to happen afterwards is not strictly dominated by some alternative future he would consider, had he chosen any other action. In other words, as every possible next vertex corresponds to one of \(i\)’s possible actions, \(i\) is rational if and only if each of these vertices is not strictly dominated, according to \(i\)’s uncertainty, by the next actual vertex (corresponding to the actual action chosen by \(i\)).

\[ \text{Rat}_i \overset{\text{def}}{=} \text{Rat}^\text{end}_i \land \text{Rat}^{\neg\text{end}}_i \]

Note that introspection on material rationality is expressed by the following valid formula in ELEG (see the syntactic proof of Lemmas A.1 and A.2 in Appendix A for details):

\[ \vdash_{\text{ELEG}} \text{Rat}_i \leftrightarrow [K_i]\text{Rat}_i \]

From the previous definition of material rationality, one can easily define Aumann’s concept of substantive rationality. In fact, according to Aumann, an agent is substantively rational if and only if “no matter where a player finds himself - at which vertex - he will not knowingly continue with a strategy that yields him less than he could have gotten with a different strategy” (Aumann [1995]). The corresponding definition in ELEG is as follows:

\[ \text{SRat}^n_i \overset{\text{def}}{=} \AX^{\leq n}\text{Rat}_i \]

\(\text{SRat}^n_i\) reads “agent \(i\) is substantively rational up to some depth \(n\).” In order to deal with perfect substantive rationality, the \(n\) parameter should not be lower than the maximum depth of the game tree.

Moreover, introspection on substantive rationality can also be deduced through Axiom \(\text{Perm}_{[K_i],\AX}\). The following formula is therefore valid in ELEG:
Theorem 3.3 For every $n \in \mathbb{N}, i \in \text{Agt}$, we have:

\[ \vdash_{\text{ELEG}} \text{S\text{Rat}}_i^n \leftrightarrow [K_i]\text{S\text{Rat}}_i^n \]

3.3 A syntactic proof of Aumann’s theorem

As already stated in the previous section, a fundamental assumption of Aumann’s theorem is that the game is in “general position”, i.e., every history of the game is associated to a unique preference value for every agent. This important notion can be defined in the logic $\text{ELEG}$ in the following way:

\[
\text{GenPos}^n \overset{\text{def}}{=} \bigwedge_{0 \leq h \leq n} \bigwedge_{k \in I, i \in \text{Agt}, \epsilon \in \text{Seq}^h} \text{AX}^{\leq n}\Box((k_i \land \langle \epsilon \rangle)\text{end}) \rightarrow \Box((\langle \epsilon \rangle)\text{end} \leftrightarrow k_i))
\]

In our syntactic proof of Aumann’s theorem we only consider game structures with uniform depth, that is, games whose end vertices have the same distance from a given vertex.

The following construction $\text{Depth}^n$ means that “the current game has a uniform depth of degree $n$ from the current vertex”. In other words, no matter what actions will be chosen in the future, an end vertex will be reached in exactly $n$ steps. This concept is thus captured by the following $\text{ELEG}$ formula:

\[
\text{Depth}^n \overset{\text{def}}{=} (\Box X)^n \text{end}
\]

This assumption, which is not stated in Aumann’s original theorem, is used here only to simplify the formal proof. One should note however that any extensive game can be represented by an extensive game with uniform depth (i.e., by adding “dummy decision nodes” where players have only one feasible action).

According to Aumann’s theorem, the following constraints must be satisfied in order for the current strategy profile to be a backward induction solution:

- the game is finite;
- the game has a uniform depth of degree $n$ from the current vertex;
- the game is in the general position;
- there is common knowledge up to level at least $n$ that at every future vertex (up to depth $n$) all agents are rational.

Theorem 3.4 For every $n, m \in \mathbb{N}$ such that $n \leq m$, we have:

\[ \vdash_{\text{ELEG}} ([\text{CK}^m_{\text{Agt}}] \bigwedge_{i \in \text{Agt}} \text{S\text{Rat}}_i^n \land \text{Depth}^n \land \text{GenPos}^n) \rightarrow \text{BI}^n \]
Note that the proof of Theorem 3.4 only requires to prove the case where \( m = n \) (see Lemma A.3 for details).

The syntactic proof of Aumann’s theorem clearly demonstrates that the temporal factor actually plays an important role in the case of reasoning in extensive form games and interpreting the backward induction. Indeed, the maximum depth of the game tree clearly justifies the level of common belief in substantive rationality required to derive the backward induction solution (i.e., \( m = n \) in Theorem 3.4). Moreover, the proof in Appendix A also shows that the hypothesis of common knowledge among all agents is not necessary to reach the backward induction. The following weaker assumption indeed suffices to derive it:

\[ \Rightarrow \] in a game of maximum depth \( n \), at every vertex where the maximum depth of the remaining subgame tree is \( m \leq n \), it is common knowledge up to some level \( m \) among players in \( J \subseteq Agt \) that all players in \( J \) are substantively rational (where \( J \) only includes those agents who may play at some future vertex).

### 3.4 A more convenient characterization of knowledge

Based on the previous statement of Aumann’s theorem, one can note from the syntactic proof in Appendix A that Theorem 3.4 can be weakened through a reinterpretation of the epistemic operator. Indeed, every proof step from Theorem 3.4 using Axiom \( T \) for the \( S5 \) knowledge operator \( [K_i] \) can still be proved using \( KD4 \) principles for the belief modal operator. Such an observation implies that a simple notion of belief (which is not necessarily truthful) is sufficient to prove Aumann’s theorem. The detailed proof of Theorem 3.4 in Appendix A shows that such a weakening of the epistemic operator is made possible mainly by Axiom \textbf{Aware} that requires agents to have introspection on their own performing action (see specific proofs of Lemmas A.1 and A.3 in Appendix A for details). In other words, agents always believe without a doubt what they’ll actually perform.

Another interesting observation from the syntactic proof of Theorem 3.4 is that agents do not require to have negative introspection over their knowledge: indeed axiom \( 5 \) from the knowledge operator \( [K_i] \) is not required to the proof. However, one should note that this axiom is required so that rationality introspection holds (see Lemma A.2 in Appendix A for details). As a direct implication, this simply means that Theorem 3.4 does not require agents to be aware of their own rationality.
In order to illustrate such an argument, let us consider a three round version of the well known centipede Game ([Rosenthal 1981]), as shown in Figure 3.2.

![Figure 3.2: The centipede Game (Alice plays at \(w_1\) and \(w_3\); Bob plays at \(w_2\))](image)

In order to formalize the game depicted in Figure 3.2 (where \((x, y)\) reads “Alice gets a payoff of \(x\) while Bob gets a payoff of \(y\)”), let us construct the following model in [ELEG].

We suppose \(Agt = \{Alice, Bob\}, Act = \{a_1, d_1, a_2, d_2, a_3, d_3\}, V = \{w_1, w_2, w_3\}, EndV = \{w_3\},\) and \(S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\};\)

\(S\) includes the eight possible strategy profiles \(s_1, \ldots, s_8\) of the centipede game, each of which includes the same three vertices from \(V\) where various actions occur:
The corresponding ELEG model is constructed as follows:

- \( \text{next}(w_1, s) = w_2 \) for \( s \in \{s_5, s_6, s_7, s_8\} \);
  \( \text{next}(w_2, s) = w_3 \) for \( s \in \{s_3, s_4, s_7, s_8\} \);

- \( Q(w_1) = Q(w_3) = \text{Alice} \);
  \( Q(w_2) = \text{Bob} \);

- \( P_{\text{Alice}}(w_1, s) = P_{\text{Bob}}(w_1, s) = 1 \) for \( s \in \{s_1, s_2, s_3, s_4\} \);
  \( P_{\text{Alice}}(w_1, s) = 0 \) and \( P_{\text{Bob}}(w_1, s) = 11 \) for \( s \in \{s_5, s_6\} \);
  \( P_{\text{Alice}}(w_1, s_7) = 16 \) and \( P_{\text{Bob}}(w_1, s_7) = 10 \);
  \( P_{\text{Alice}}(w_1, s_8) = P_{\text{Bob}}(w_1, s_8) = 15 \);
  \( P_{\text{Alice}}(w_2, s) = 0 \) and \( P_{\text{Bob}}(w_2, s) = 11 \) for \( s \in \{s_1, s_2, s_5, s_6\} \);
  \( P_{\text{Alice}}(w_2, s) = 16 \) and \( P_{\text{Bob}}(w_2, s) = 10 \) for \( s \in \{s_3, s_7\} \);
  \( P_{\text{Alice}}(w_2, s) = P_{\text{Bob}}(w_2, s) = 15 \) for \( s \in \{s_4, s_8\} \);
  \( P_{\text{Alice}}(w_3, s) = 16 \) and \( P_{\text{Bob}}(w_3, s) = 10 \) for \( s \in \{s_1, s_3, s_5, s_7\} \);
  \( P_{\text{Alice}}(w_3, s) = P_{\text{Bob}}(w_3, s) = 15 \) for \( s \in \{s_2, s_4, s_6, s_8\} \);

As shown in Halpern [2001], Stalnaker illustrates his argument by extending Aumann’s model with a selection function that characterizes a notion of “closeness” between two strategy profiles for every given vertex.

This corresponds to extending an extensive game model in ELEG with the following function \( F \):

- \( F \) is a total function \( F : S \times W \rightarrow S \) mapping every strategy profile and vertex to a strategy profile such that:
  
  \( F_1 \) if \( F(s, w) = s' \), then for every \( v \in V \) s.t. \( wR^*v, s(v) = s(v)' \);
  
  \( F_2 \) if \( F(s, w) = s' \), and \( vR^*w \), and \( v \neq w \), then \( \exists v' \in V \) s.t. \( \text{next}(v, s') = v' \) and \( v'R^*w \).

(Where \( R^* \) denotes the reflexive transitive closure of \( R \))

In other words, given \( s, s' \in S \) and \( w \in W \), \( F(w, s) = s' \) if and only if \( s \) and \( s' \) specify the same actions at every vertex that can be reached from \( w \). Applied to the above centipede game, \( F \) can partially be defined as follows:
Thus, according to Stalnaker, an agent $i$ is rational at some vertex $v$ along some strategy profile $s$ if $i$ is rational at $v$ in $F(s, v)$. Following this interpretation, it is easy to check that common belief in substantive rationality (according to the above rule) does not imply the backward induction outcome (see Halpern [2001] for a detailed analysis).

However, while Stalnaker’s concept of substantive rationality is more realistic than Aumann’s, the main remaining weakness carried out in both models concerns the structure of the epistemic relation that only considers states/strategy profiles. Indeed, both models restrict agents to have the same epistemic state at every vertex from the same strategy profile (i.e., agents have the same uncertainty regarding strategy profiles no matter which vertex they are in). Such a constraint can be observed through the semantic definition of $E_i$ in ELEG. Moreover, Axiom $Perm_{[K_i],AX}$ reflects the unrealistic structure of such epistemic relations. Obviously Axiom $Perm_{[K_i],AX}$ is very strong as it assumes that players know at the beginning of the game what they will do at any reachable state in the future where they have to play. This simply means that players can neither learn nor forget through the gameplay. In order to allow the players to revise their beliefs and act more realistically as they advance through the game, one therefore needs to consider vertices along with strategy profiles within the epistemic relation. In this way, a player who finds out that a possible strategy is discarded by another’s move at some vertex may then update his knowledge, allowing him to later act accordingly.

However, unlike in Aumann and Stalnaker’s models discussed above, the bidimensional semantics of ELEG allows such a revision: indeed, every ELEG formula is interpreted with respect to a vertex/strategy profile pair. Hence, we can here interpret the epistemic modal operator by means of an equivalence epistemic relation $E_w^i$ on strategy profiles from $S$ for every agent $i \in \text{Agt}$ and every vertex $w \in V$. Consequently, agents’ uncertainty over strategy profiles can evolve through time.

Given this change on the epistemic relations, the truth condition of the knowledge operator then becomes:

- $M, w, s \models [K_i] \varphi$ iff $M, w, s' \models \varphi$ for all $s'$ such that $sE_w^i s'$.

Considering these new epistemic relations $E_w^i$, the previous Constraint $C3$ has to be reformulated as follows:

$$C3^* \quad \text{if } sE_w^i s' \text{ and } \Omega(w) = i, \text{ then } s(w) = s(w)'$$

Moreover, the following constraints need to be introduced in order to keep Axiom $Perm_{[K_i],AX}$ as in Table 3.1:
According to Constraint C6, agents will never forget their current uncertainty over strategy profiles in every reachable vertex. In other words, C6 simply means that agents will always have a perfect recall of their past uncertainty throughout the game. According to Constraint C7, agents are always aware of their future uncertainty over strategy profiles in every reachable vertex. In other words, C7 means that agents will never be able to discard strategy profiles and therefore learn as they advance through time.

Let us provide the axiom corresponding to Constraint C6 alone (without Constraint C7):

\[
\text{Perm}^{\ast}_{[K_i],AX} \quad [K_i]AX \phi \rightarrow AX[K_i] \phi
\]

Note that Axiom \text{Perm}^{\ast}_{[K_i],AX} is simply a weaker version of the initial Axiom \text{Perm}_{[K_i],AX} from Table 3.1. It is clearly showed in Appendix A that Constraint C6 along with its corresponding Axiom \text{Perm}^{\ast}_{[K_i],AX} are sufficient to the syntactic proof of Theorem 3.4. Such an observation simply implies that Aumann’s theorem holds even though agents are learning through the game (i.e., Constraint C7, which is not necessary, can be removed). As another interesting consequence of removing Constraint C7, agents may not even be aware of their own future moves (Constraint C3$^*$ only requires them to be aware of their current move).

However, it is worth noting that the following intuitive condition should hold, independently of previous constraints C6 and C7:

\[
\text{C8} \quad \text{if } \mathcal{Q}(w) = i \text{ and } w \mathcal{R}v, \text{ then } s \mathcal{E}_i w s' \text{ iff } s \mathcal{E}_i s'
\]

According to Constraint C8, an agent’s action cannot affect his epistemic state: one’s uncertainty should remain the same before and after performing a given action.

The axiom corresponding to Constraint C8 in ELEG then becomes as follows:

\[
\text{Perm}^{**}_{[K_i],AX} \quad \text{turn}_i \rightarrow ([K_i]AX \phi \leftrightarrow AX[K_i] \phi)
\]

As for Axiom \text{Perm}^{**}_{[K_i],AX}, Axiom \text{Perm}^{\ast}_{[K_i],AX} is also a weaker version of the initial Axiom \text{Perm}_{[K_i],AX} from Table 3.1.

In order to demonstrate that Constraint C6 is necessary and Constraint C8 is not sufficient to the proof of Theorem 3.4, let us construct an epistemic model $M$ for the above centipede game such that Constraint C8 holds while both constraints C6 and C7 are removed. Figure 3.3 illustrates the epistemic relations in $M$ where reflexive relations are implicit for every strategy profile at all vertices, and $F$ denotes Stalnaker’s “closeness” function presented above.
\(E_w^{Alice}(s) = E_w^{Bob}(s) = \{s_5, s_7\} \) for every \(w \in \{w_1, w_2, w_3\}\);
\(E_w^{Alice}(s) = \{s\} \) for every \(s \in S\setminus\{s_5, s_7\}\) and every \(w \in \{w_1, w_2, w_3\}\);
\(E_w^{Bob}(s) = \{s\} \) for every \(s \in S\setminus\{s_7, s_8\}\);
\(E_w^{Bob}(s) = \{s\} \) for every \(w \in \{w_2, w_3\}\) and \(s \in \{s_1, s_2, s_5, s_6\}\);

In the above model, let us assume that \(s_3\) is the actual strategy profile. At \(w_1\) in \(s_3\) (see Figure 3.3(a)), both players know what would happen at every possible future vertex (i.e., at \(w_1\) in \(s_3\), each player knows that Alice plays \(d_1\) at \(w_1\), Bob would play \(a_2\) at \(w_2\), and Alice would play \(d_3\) at \(w_3\)). However, at the unexpected vertex \(w_2\) in \(s_3\) (see Figure 3.3(b)), Bob somehow becomes uncertain about Alice’s move at \(w_3\). It is easy to show that Bob is rational at \(w_2\) in \(s_3\), and consequently Bob is substantively rational at \(w_1\) in \(s_3\). Similarly, as Alice is rational to play \(d_1\) at \(w_1\) (see Figure 3.3(b)) and also to play \(d_3\) at \(w_3\) (see Figure 3.3(c)), she is also substantively rational at \(w_1\) in \(s_3\). Therefore, this implies that common knowledge of substantive rationality holds at \(w_1\) in \(s_3\). Following the same reasoning, it can easily be shown that common knowledge of substantive rationality also holds at
$w_1$ in $s_1$. Note that common knowledge of Stalnaker's substantive rationality (see above) similarly holds for both $s_1$ and $s_3$ at $w_1$.

However, performing a game theoretic analysis of the above centipede game leads to the following unique backward induction solution: Alice plays strategy $(d_1, d_3)$ while Bob plays strategy $(d_2)$. One should note that $s_1$ is indeed the only strategy profile that fits the backward induction restriction because the game is in general position (i.e., with different payoffs at all leaves for every player). The above epistemic model therefore illustrates that common knowledge of rationality does not necessarily imply backward induction.

Therefore, this analysis indicates that Constraint C6 and corresponding Axiom \( \text{Perm}^{[K_i]AX} \) are necessary to allow common knowledge of substantive rationality to derive backward induction. In other words, agents are simply required to have perfect recall through the game in order to follow this prediction. However, such a constraint clearly remains very strong since it implies that, even in the case of some unexpected event occurring, an agent would still know what he used to know before such an event. In order to investigate this issue, let us look at some possible revisions of Bob’s epistemic states that can be made at vertex $w_2$ in model $M$ from Figure 3.3 where such perfect recall is removed.

According to model $M$, at $w_1$ in $s_3$, Bob knows that Alice is substantively rational, and if $w_2$ was actually reached, then he would not know that Alice is rational. Note that, as suggested by Stalnaker, Bob’s epistemic state at $w_2$ in $s_3$ should be evaluated in the “closest” strategy profile $F(w_2, s_3) = s_7$.

\[
M, w_1, s_3 \models [K_{Bob}]\text{SRat}_2 Alice
\]

and

\[
M, w_2, s_7 \models \neg [K_{Bob}]\text{SRat}_1 Alice
\]

Such a plausible interpretation can easily be justified as follows: upon reaching the unexpected vertex $w_2$, Bob (who is assumed to be rational) then revises his beliefs so that Alice may not be rational in the future because she was not rational to previously play $a_1$ at $w_1$. In fact, it is straightforward to show that, for any setting of the epistemic relations $E_{Bob}$ and $E_{Alice}$ in some alternative model $M'$, Bob’s substantive rationality cannot be compatible with Bob’s knowledge of Alice’s substantive rationality at $w_1$ in $s_7$.

\[
M', w_1, s_7 \not\models \text{SRat}_2 Bob \land [K_{Bob}]\text{SRat}_2 Alice
\]

The main consequence of this revision of Bob’s epistemic states is that he can then be rational to select $a_2$ at $w_2$ in $s_3$, because he considers it possible that Alice will play $a_3$ at $w_3$. However, one may wonder whether this is the best interpretation Bob should follow. In fact, Bob may instead ask himself whether Alice could have
acted rationally by selecting $a_1$ at $w_1$. In order to answer this question, let us now look at strategy profile $s_1$ in $M$.

According to model $M$, at $w_2$ in $s_1$, Bob still knows that Alice is substantively rational, and he would maintain this knowledge if $w_2$ was actually reached (i.e., in the “closest” strategy profile $F(w_2, s_1) = s_5$). In this case, he would however revise his beliefs about Alice’s epistemic state as follows:

$$M, w_1, s_1 \models [K_{Bob}](SRat^2_{Alice} \land [K_{Alice}]SRat^2_{Bob} \land [K_{Alice}][K_{Bob}]SRat^2_{Alice})$$

and

$$M, w_2, s_5 \models [K_{Bob}](SRat^4_{Alice} \land [K_{Alice}]SRat^2_{Bob} \land \neg[K_{Alice}][K_{Bob}]SRat^4_{Alice})$$

In other words, this interpretation simply means that, if vertex $w_2$ is actually reached, then Bob would simply learn that Alice has limited knowledge (e.g., she may have limited cognitive abilities that prevent her from knowing that Bob knows that she is rational). In fact, unlike strategy profile $s_7$, it is easy to show that, at $w_1$ in $s_5$, Bob’s rationality can be compatible with Bob’s knowledge that (1) Alice is rational, and that (2) Alice also knows that he is rational.

$$M, w_1, s_5 \models SRat^2_{Bob} \land [K_{Bob}](SRat^2_{Alice} \land [K_{Alice}]SRat^2_{Bob})$$

However, for any setting of the epistemic relations $E_{Bob}$ and $E_{Alice}$ in some alternative model $M''$, in $s_7$, Bob cannot be rational and simultaneously know that (1) Alice is rational, that (2) Alice knows that he is rational, and that (3) Alice knows that he knows that she is rational.

$$M'', w_1, s_5 \not\models SRat^2_{Bob} \land [K_{Bob}](SRat^2_{Alice} \land [K_{Alice}]SRat^2_{Bob} \land [K_{Alice}][K_{Bob}]SRat^2_{Alice})$$

As a result, it appears that the current interpretation of Bob’s epistemic state does not conflict with the prediction of the backward induction: as a consequence of the players’ epistemic states at $w_2$ in $s_5$ from $M$ (see Figure 3.3(b)), Bob should indeed select $d_2$ at $w_2$ as the unique best response to knowing that Alice’s only rational move at $w_3$ is $d_3$.

Following this analysis, although both interpretations appear to be intuitive and plausible, one might wonder how Bob should actually revise his beliefs at vertex $w_2$: should he believe that Alice was not rational to play $a_1$ at $w_1$ (as in $s_3$ from $M$)? Or should he instead believe that she is rational with some limited cognitive abilities (as in $s_1$ from $M$)? Clearly, giving up the belief that Alice is rational would be very naive. It is indeed in Bob’s best interest to continue believing in Alice’s rationality at $w_2$: by giving up his belief about Alice’s rationality,
Bob would then expose himself to the risk of being exploited by Alice at $w_3$ if he appears to be wrong (i.e., if he plays $a_2$ at $w_2$ and Alice then plays $d_3$ at $w_3$, as in $s_3$).

More generally, while this analysis emphasizes the relevance of Stalnaker’s criticism about Aumann’s concept of substantive rationality, it also indicates that, in the context of perfect information games that are in the general position, revising epistemic states at unexpected vertices should still lead to the backward induction solution. In fact, although perfect recall clearly remains a strong assumption (cf. Constraint $C6$), the unrealistic beliefs resulting from it do not appear to affect the predicted solution in any way: in the above centipede game, whether Bob maintains his belief about the fact that Alice believes that he believes in her rationality is irrelevant to determine Bob’s rational choice at $w_2$ (either way, he should rationally play $d_2$ at $w_2$, as a result of still believing that Alice is rational).

However, this analysis points out to the main actual criticism of Aumann’s concept of substantive rationality in the context of perfect information games, which can be illustrated through a variant to the above centipede game, as depicted in Figure 3.4.

![Figure 3.4: A variant of the centipede Game (Alice plays at $w_1$ and $w_3$; Bob plays at $w_2$)](image)

The only difference between the game from Figure 3.4, and the previous game from Figure 3.2 is that Alice gets to earn a payoff of 17 instead of 1 by playing $d_1$ at $w_1$ (every other payoff remains unchanged). In this case, the backward induction solution remains as in the game from Figure 3.2: Alice will play strategy $(d_1, d_3)$ while Bob will play $(d_2)$. However, the particular property of this new version of the game is that, no matter what would actually happen in the subgame (at $w_2$ and $w_3$), playing $d_1$ at $w_1$ is always better for Alice. As a consequence, while playing $d_1$ at $w_1$ remains optimal for Alice, one may wonder whether the backward induction prediction in the subgame is reasonable. Indeed, if vertex $w_2$ were to be actually reached, then Bob would learn that Alice did not act rationally at $w_1$. 

---

68
In this case, Bob may reasonably infer that, since Alice was irrational once, it is possible that she will be again irrational in the future (i.e., at $w_3$). Therefore, while it remains rational for Bob to play $d_2$ at $w_2$, it now also becomes rational for him to play $a_2$ (in response to his expectation that Alice may possibly play $a_3$ at $w_3$). As a result, Aumann’s concept of substantive rationality (i.e., rationality that holds at all vertices of the game) clearly appears to be counter-intuitive in such games: rationality in vertices that cannot be reached rationally does indeed not make any sense. On the other hand, note that Stalnaker’s more realistic definition of substantive rationality (through the “closeness” function $F$, as shown above) cannot be expressed in such a game. As a result, one can therefore claim that, in order for Aumann’s concept of substantive rationality to make sense, it should strictly be applied to extensive games where every vertex can be reached rationally (which then rules out the game from Figure 3.4).

However, while both Aumann and Stalnaker’s models do not allow for the detailed epistemic analysis of extensive games presented here because they do not incorporate time in their definition of the epistemic relations, it is worth noting that our logical framework can be seen as an alternative approach to other relevant economic models such as Battigalli and Siniscalchi [1999, 2002]), which presents an epistemic model that also explicitly captures the temporal dimension of extensive games and allows to reason about counterfactuals. More generally, the above analysis demonstrates the importance of considering the temporal factor in extensive games, which can only offer a more realistic interpretation of the reasoning process followed by boundedly rational individuals.

### 3.6 Related work in logic

We are not the first to provide a logical analysis of extensive games. Several logical systems have been proposed which support reasoning about this class of games. We here discuss some of these systems and compare them with our logic ELEG.

In van Benthem [2002], van Benthem analyzes extensive games using different modal languages such as propositional dynamic logic (PDL), PDL with converse, and a modal forcing language which allows to express what a player can bring about in a given extensive game, no matter what the other players do. Moreover, he also studies a variety of notions of game equivalence based on the notion of bisimulation. Although van Benthem shows how PDL extended with epistemic operators can represent extensive games with imperfect information, he does not consider the concept of rationality which is a fundamental element of Aumann’s epistemic analysis of extensive games. It is worth noting that, differently from our logic ELEG, standard PDL would fail to define such a concept, because it can neither identify the current strategy that is going to be played nor express
what will be true at every possible next vertex along the current strategy profile (which is done through the operator $AX$ in ELEG). Moreover, our logic ELEG shows that defining strategy profiles explicitly in the object language — as done in PDL — is not necessary to express interesting game-theoretic concepts such as rationality and backward induction.

Related to van Benthem’s work is Ramanujan & Simon’s work (Ramanujam and Simon [2008a,b]) who have recently proposed an elegant approach to extensive games based on dynamic logic. However, Ramanujan & Simon do not deal with epistemic aspects of extensive games, as their logic does not have operators for representing epistemic states of players. The game logic presented in Parikh [1985] also lacks epistemic operators, therefore preventing a formalization of the concept of epistemic rationality and a logical analysis of Aumann’s theorem. Bonanno’s logical account of extensive games (Bonanno [2001, 2002]) has the same limitation. He proposes a variant of dynamic logic extended with temporal operators for (branching) future and (linear) past and shows how his logic can be used to characterize the solution concept of backward induction. But, like Ramanujan & Simon’s logic, Bonanno’s logic does not have epistemic operators which are required to represent Aumann’s notion of rationality and the statement of Aumann’s theorem. The same remark also applies to some recent work (Surowik [2004]), which presents a similar logical approach to extensive games without considering the epistemic aspects.

ATL-based approaches to extensive games presented in Walther et al. [2007] and van Der Hoek et al. [2005] come closer to our current approach. For instance, in Walther et al. [2007] a variant of ATL (Alternating-time temporal logic) with explicit strategies called ATLES (Alternating-time logic with explicit strategies) is proposed which allows to define solution concepts such as backward induction. The interesting aspect of ATLES, compared to ATL, is that one can explicitly reason about strategies in the object language. However, like Ramanujan & Simon and Bonanno, ATLES misses epistemic operators necessary to define Aumann’s notion of rationality. Another important difference between ATLES and our logic ELEG is that in ATLES formulas are interpreted with respect to states, whereas in ELEG they are interpreted with respect to state/strategy profile pairs (in this sense, ELEG semantics is bidimensional). The latter is an advantage because, differently from ATLES, it is possible in ELEG to reason about what will be true at every possible next vertex along the current strategy profile. We have shown that this is fundamental for expressing in the object language Aumann’s notion of rationality and the statement of Aumann’s theorem.

Bonanno’s logic has four kinds of operators for past and future describing: (1) what is going to be the case at every future vertex of the game tree, (2) what has always been the case at every past vertex, (3) what is going to be the case at every predicted future vertex of the game tree, and (4) what has always been the case at every past vertex at which today was predicted.
In Vestergaard et al. [2006], the authors propose an alternative way of proving Aumann’s theorem by using a purely proof-theoretic approach based on type theory. Differently from Vestergaard et al.’s approach, our approach based on modal logic has the advantage of combining a proof-theoretic analysis of extensive games — which is what we have done in Section 3.3 — with a model-theoretic semantics.

3.7 Conclusion

In this chapter, we have introduced a logical framework that provides an alternative way of representing extensive form games as compared to their usual specification in economics.

We showed that our logic is sufficiently general for our purpose to reason about dynamic epistemic games, as illustrated by the well known concepts of rationality and backward induction. Although these concepts have been extensively studied in economics, very few logical analyses have been proposed up to now. While several related work discuss and present some logical approaches to epistemic reasoning in such extensive games, none of these define a logic expressive enough to represent syntactically both the epistemic concepts and the equilibrium solutions. By the formal syntactic proof of Aumann’s theorem, we demonstrate that our logic is capable to fill this gap and provide further interesting information about those concepts.

However, it is worth noting that, even though such an epistemic analysis provides a very intuitive justification for why individuals often deviate from the optimal equilibrium solutions in social interactions, it does not always suffice. As an example, various experimental studies of the well known trust game (as it is depicted in Figure 3.1 from Section 3.1.2) have indicated that people often cooperate with one another in this situation (i.e., they reach the (3,3) outcome in Figure 3.1). Yet, performing an epistemic analysis of this game suggests that the second player can never be rational to cooperate, no matter his belief state. Therefore, instead of considering such a behavior as irrational, we will argue, through the next chapters, that one’s rationality needs to be revised by taking into account important social factors such as one’s social connections with other individuals involved in the interaction.
Chapter 4

Social Ties and Strategic Coordination

“We do not succeed in changing things according to our desire, but gradually our desire changes.”
— Marcel Proust
The Sweet Cheat Gone (1925)

“The primal scene of morality... is not one in which I do something to you or you do something to me, but one in which we do something together.”
— Christine Korsgaard
The Reasons We Can Share (1993)

In classical economic theories, most models assume that agents are self-regarding and maximize their own material payoffs. However, as already mentioned in Chapter 2, important experimental evidence from economics and psychology have shown some persistent deviation from such individualistic behavior in many strategic situations. These results suggest the need to incorporate social preferences into game theoretic models. Such preferences describe the fact that a given player not only considers his own material payoffs but also those of other players (Margolis [1982]). The various social norms created by the cultural environment in which human beings live give some ideas of how such experimental data could be interpreted: fairness, inequity aversion, reciprocity and social welfare maximization are concepts that behavioral economists are familiar with, and which have been shown to play an important role in interactive decision making (e.g., see Charness and Rabin [2002]; Fehr and Schmidt [1999]; Rabin [1993a]).
In fact, various simple economic games, such as the trust game (Berg et al. [1995]) and the ultimatum game (Güth et al. [1982]), have been extensively studied in the past years because they illustrate well the weakness of traditional game theory and its assumption of individualistic rationality. Moreover, given the little complexity carried out in such games, the bounded rationality argument (Gigerenzer and Selten [2001]) does not seem sufficient to justify observed behavior. Social preferences appear as a more realistic option because they allow to explain the resulting behaviors while still considering rational agents.

However, although many economic experimental studies (e.g., Berg et al. [1995]; Güth et al. [1982]) have shown that people genuinely exhibit other-regarding preferences when interacting with perfect strangers, one may wonder to what extent the existence of some social relationships between individuals may influence behavior. Throughout this dissertation, we refer to such social relationships as ‘social ties’. Indeed the dynamic aspect of social preferences seems closely related to that of social ties: one may cooperate more with a friend than with a stranger, and doing so may eventually enforce the level of friendship.

Our attempt, through this chapter, is to study the possible effects that positive social ties can have on human cooperation and coordination. Our main hypothesis is that such relationships can influence a player’s choice by modifying his preferences: an agent may choose to be fair conditionally on the relative closeness to his partner(s). In order to investigate these questions, we propose a theoretical analysis of a new kind of two player game that allows us to disentangle predictions from theories based on self-interest, social preferences, and social ties. Furthermore, we demonstrate the need to introduce an alternative model to capture the concept of social ties as continuous variables. Indeed, while we claim that social ties strongly rely on group identification, we show that considering the concept of team reasoning is too limited to fill this purpose as it is built upon a binary interpretation of group identification (i.e., either one identifies with a group or not).

4.1 A definition of social ties

No formal definition of a social tie is provided either in the literature on social psychology or in the experimental economics literature focused on social preferences. Thus, given the vagueness and the ambiguity that the term may suggest, we begin by clarifying the concept that we consider.

First, we choose to restrict our study only to those ties that can be judged to be positive: examples include relationships between close friends, married couples, family relatives, classmates, etc. In contrast, negative ties may include relationships between people with different tastes, from different political orientations, with different religious beliefs, etc.
In order to specify the foundations of such social ties and the possible reasons for their emergence, let us consider the well-known concept of social identity from social psychology. According to social identity theory (Hogg [2002]; Tajfel and Turner [1979]), an individual’s social identity is built upon a set of social features, each of which may refer to any type of salient characteristics that can be shared by individuals in a particular context. For example, a person may identify himself as a student of the university of Toulouse, a supporter of Barcelona’s soccer team, a Democrat, a Catholic, etc.

According to various theories in social psychology (see, e.g., Abrams and Hogg [2006]; Hogg [2000]), the construction of an individual’s social identity is determined by two complementary motivations. The first motivation is self-enhancement, which is underpinned by one’s individual need to promote self-esteem (as pointed out by Luhtanen and Crocker (Luhtanen and Crocker [1992]), “Being a member of a social group is an important reflection of who I am”). Reduction of subjective uncertainty about one’s perceptions, attitudes, feelings, behavior, and one’s self-concept and place within the social world is the second motivation.

It can be reasonably assumed that people can give different degrees of importance to those social features defining their social identity, depending on the context: for example, while one’s identification as a soccer player is more important than one’s identification as a student during a soccer game, the reverse may hold for the same individual during a math exam at the university.

Following this interpretation, our claim is that:

**Statement 4.1.0.1** A social tie between two individuals exists if and only if they share the same social features defining their social identities, and this is common belief among them.

Note that the previous claim implies that a social tie is necessarily bilateral in the sense that, if an individual \(i\) is tied with another individual \(j\), then \(j\) is also tied with \(i\). For example, an individual who believes to share the same political convictions with a given politician cannot induce a social tie as long as the latter does not also believe so (one could speak of the existence of a unilateral tie in this case, though it is not “social” according to the above statement).

Moreover, the previous statement simply characterizes the minimal condition for the existence of a social tie. As an illustration, one can consider the well known Minimal Group Paradigm (MGP) from Tajfel [1970], which corresponds to an experimental methodology from social psychology that investigates the minimal conditions required for discrimination to occur between groups. Experiments using this approach (Tajfel et al. [1971]) have revealed that arbitrary and virtually meaningless distinctions between groups (e.g., the colour of their shirts) can trigger a tendency to cooperate more with individuals within one’s own group than with others. In this case, one should note that such meaningless social features
satisfy the minimal condition for being considered as a social tie from the previous statement. However, in principle such social tie should be quite weak.

In this respect, it is worth mentioning that an important property of social ties lies in its quantitative aspect, that is, two individuals can be more or less socially tied with each other. To be more precise, we assume that a social tie between two individuals can be measured on a scale ranging from 0 to 1, where 0 and 1 respectively stand for the minimum and maximum strength for the tie.

This interpretation therefore suggests that the strength of a social tie can be determined by the \textbf{quantity and importance of shared social features}. One can indeed assume that sharing a high number of social features (defining one’s social identity) with high importance leads to a high social tie value. On the other hand, having conflicting social characteristics, or sharing a low number of features with high importance, or sharing a high number of features with low importance can lead to a lower tie value.

Moreover, another aspect that, we believe, influences the strength of a social tie between two individuals is the \textbf{quantity and quality of past interactions} between them. More precisely, given two individuals sharing a certain number of social features with a given importance, the strength of the tie between them is higher in the situation in which the two individuals had frequent meaningful interactions in the past than in the situation in which there were no previous meaningful interactions\footnote{With the term “meaningful” we mean that during past interactions, the two individuals had the occasion to know each other by exchanging ideas, opinions, sharing positive emotions (e.g., they mutually enjoyed playing tennis together), etc.}.

As a concrete example to illustrate the previous interpretation, one may consider the case of online dating systems on the internet. Those systems, which are clearly meant to build social ties between individuals (assuming an affective tie is a special case of a social tie), are based on the matching of social features that define their social identities. However, while one cannot deny the effectiveness of such systems (Hitsch et al. [2010]), it is suggested in Frost et al. [2008] that some interaction between two individuals is also important as it can allow them to know each other more accurately. Indeed, by providing a way to obtain reliable information about one another, social interactions happen to be a relevant tool against possibly inaccurate stereotypes, which can often be considered as an unfortunate consequence of categorizing individuals into social groups, as implied by social identity theory.

The following points summarize our interpretation of social ties:

- The minimal criterion for the existence of a social tie between two individuals is for them to commonly believe that they share the same social features that define their social identities.
A social tie between two individuals has a quantitative dimension which depends on the following variables:

1. The quantity and importance of shared social features that define both individuals’ social identities.
2. The quantity and quality of past interactions between both individuals.

Following our interpretation, one might then argue that the situation described by the minimal group paradigm (MGT) satisfies the minimal condition for the existence of a social tie, even though this tie has a relatively low degree of strength (the number of shared social features is one) and its importance might be considered to be reasonably low.

### 4.2 A theory of how to model social ties

In this section, we introduce a novel model that characterizes the agents’ behavior in the presence of social ties. Our model of social ties shares features with both team reasoning and social preferences theories.

Similarly to team reasoning theories, our model is built on the concept of group identification, which is of high relevance when considering social ties. In fact, individuals that are socially connected may be expected to identify themselves with the same group, which may consequently lead them to choose actions as a member of this group. In section 4.7 we discuss various theories of team reasoning, at the same time indicating which properties are in common with our approach to social ties. In section 4.8 we underline the inadequacy of such theories to interpret in full the effects of social ties, thereby claiming for a novel approach – the one introduced in this section – able to capture specific key features of social ties left aside by team reasoning.

Similarly to theories of social preferences, our starting assumption is that a social tie between two individuals induces them to behave according to some aggregation of their individual preferences. In section 4.5 we consider two leading theories of social preferences (inequity aversion and fairness) that are easily comparable to our approach in the specific interactive strategic situation where the behavioral effects of social ties are evaluated.

More precisely, our current approach is inspired by the existing concept of empathetic preferences as presented by Binmore in Binmore [2005]: an agent’s empathetic preferences consist in combining his actual own preferences with his preferences when imagining himself in the other agent’s position. In other words, an empathetic agent does not take into account the other’s actual decision, he instead only reasons about his own decision in the other’s position (taking into
account the other’s preferences). This concept also refers to the existence of a “veil of ignorance”, as introduced by Rawls in Rawls [1971], behind which agents make their decision without knowing in which player’s position they will actually act. As a consequence, such an empathetic behavior can be reduced to simply choosing the corresponding action from the strategy profile that maximizes the group utility.

One should note that our model is also related with Alger and Weibull’s model of Homo Moralis (Alger and Weibull [2012]). The difference is that their model requires the game to be symmetric\(^1\), whereas our model does not have any restriction and can apply to all sorts of games.

Formally, let us consider two players \(i\) and \(j\). Moreover, let \(S_i\) and \(S_j\) respectively denote the set of \(i\)’s strategies and the set of \(j\)’s strategies, and \(\pi_i(s_i, s_j)\) the material payoff function for player \(i\) when both \(i\) and \(j\) respectively play their strategy \(s_i\) and \(s_j\). For every \(s_i \in S_i\) and \(s_j \in S_j\), the Social Ties utility function of player \(i\) is given by:

\[
U_{ST}^i(s_i, s_j) = (1 - k_{ij}) \cdot \pi_i(s_i, s_j) + k_{ij} \cdot \max_{s_j' \in S_j} U(s_i, s_j')
\]

where \(k_{ij} \in [0, 1]\).

The function \(U(s_i, s_j)\) stands for the group utility function, which may be characterized by one of the following two well-known principles.

Let us first define a group utility function \(U_m(s_i, s_j)\) that satisfies Rawls’ max-min criterion (Rawls [1971]).

\[
U_m(s_i, s_j) = \min\{\pi_i(s_i, s_j), \pi_j(s_i, s_j)\}
\]

This criterion corresponds to giving infinitely greater weight to the benefits of the worse-off person.

As an alternative, one may also consider a function of social welfare \(U_s(s_i, s_j)\) that satisfies classical utilitarianism (i.e., by maximizing the total combined payoff of all players).

\[
U_s(s_i, s_j) = \pi_i(s_i, s_j) + \pi_j(s_i, s_j)
\]

Parameter \(k_{ij}\) in the Social Ties utility function measures agent \(i\)’s subjective social tie towards agent \(j\). Setting \(k_{ij}\) to 0 corresponds to a non-existing tie (e.g., \(j\) is a perfect stranger to \(i\)) whereas setting \(k_{ij}\) to 1 means that \(i\) feels socially very close to \(j\) (e.g., \(j\) is \(i\)’s best friend). In the latter case, one should note that, in the presence of a strong tie with agent \(j\), agent \(i\) does not face a strategic problem anymore: indeed, \(j\)’s strategy \(s_j\) becomes irrelevant to the calculation of \(i\)’s utility.

\(^1\)A game is symmetric when all players can switch roles without changing their strategies and the associated payoff.
Thus, agent $i$ only needs to solve a classical problem of individual decision making by selecting the action from the strategy profile which maximizes the group utility. As a result, $i$’s action may then be interpreted as “doing the right thing for the group assuming that all other players also do the right thing for the group”.

In order to make our approach operational, through the next section we propose a concrete scenario involving strategic interaction among players. The specific game that we consider leads to theoretical predictions under our social ties model that differ both from those obtained under the traditional assumption of self-interested players and by those provided by well-known theories of social preferences. Hence, we think that this example is appropriate to show how our novel approach can explain players’ behaviors not captured by leading approaches respectively in traditional and behavioral game theory.

4.3 A coordination game with outside option

Having previously analysed the main characteristics of social ties, we now propose the following game that appears to be well suited to study their behavioral effects.

Two colleagues, Alice and Bob, have agreed to cook together and eat in Alice’s place, though Bob likes cooking less than Alice. Half an hour before the dinner, Alice makes a phone call to Bob so as to express her interest to go out to her favorite Japanese restaurant, whereas Bob instead suggests to go to a nice Italian restaurant that he recently heard about, with the two restaurants located on opposite sides of the city. Alice then makes it clear that she prefers them cooking rather than Italian food. On the other hand, Bob indicates that, even though he does not particularly enjoy cooking, he still prefers it to the Japanese cuisine. Moreover, Bob reminds Alice of his seafood allergy, which makes eating Japanese more costly for him than eating Italian is costly for Alice. Both individuals also commonly realize that, as the restaurants in question are very popular, they are always overcrowded, which makes it impossible for a person alone to get a table there. Finally, Alice notifies Bob of her intention to go eat outside before the phone call suddenly gets interrupted.

In this scenario, the main questions that arise are the following: assuming Alice and Bob barely know each other, will they manage to meet and eat at the same restaurant that evening? Would they behave differently if they were very close friends instead?

For example, the fact itself that Alice would notify Bob her intention to go eat outside may depend on their social tie. And even the sign of this dependence cannot be easily disclosed through intuition alone. Indeed, on the one hand, a close relationship with Bob should lead Alice to take the “risk” of leaving her place so as to possibly meet him in his preferred restaurant, thereby maximizing their group
utility function. On the other hand, the social tie might lead Alice to confirm that they will both eat at her place before the unexpected interruption of the phone call: if she knows that Bob, because of their social tie, is “uncertain” between maximizing the group utility function (going to his preferred restaurant) and behaving in a fully rational way (going to her preferred restaurant), then she should safely reaffirm that they eat at her place, in order to prevent miscoordination.

In order to formally analyze the dilemma provided by this situation, let us define the corresponding abstract representation, as shown in Figure 4.1, which we will call the Entrance game. This Entrance game defines a two player game organized in two stages. Throughout this chapter, we denote Player $a$ as Alice and Player $b$ as Bob. During the first stage, only Alice is active, and she has to choose between $In$ or $Out$ (i.e., whether to go eating outside or stay at Alice’s place). In the latter case, the outcome is more beneficial to Alice who gets 20 (Bob only gets 10 as it is more costly for him to cook than it is for Alice). On the other hand, if Alice chooses to go eat outside (i.e., $In$), both players enter the second stage of the game that corresponds to a basic coordination game. $A$ defines either Alice or Bob’s choice to go to the Japanese restaurant. Similarly, $B$ defines either Alice or Bob’s choice to go to the Italian restaurant. If both coordinate on the $(A, A)$ solution, then Alice and Bob get 35 and 5, respectively, and if both coordinate on the $(B, B)$ solution, then they get 15 (Alice) and 35 (Bob). In any other case (i.e., where they choose different restaurants and are therefore unable to get a table), both players win nothing (0).

As we will detail in the next sections, this game provides an ideal environment to discriminate between different theories of how two socially tied people should act.

One may note that our Entrance game corresponds to a variant of the Battle of the Sexes (BoS) game with outside option (see Cooper et al. [1993]). Indeed, as shown in Figure 4.2(a), the only difference lies on the symmetrical property within the coordination subgame that we voluntarily removed here: unlike in the BoS game, the lowest payoff is different in the two coordination outcomes (eating Japanese is more costly to Bob than eating Italian is for Alice: $5 \neq 15$). The main motivation to introduce this type of asymmetry is to create some incentives for players to favour the group as a whole (in fact, there is no unique best outcome for the group in a BoS-like subgame). However, one may note that the dilemma introduced by the Entrance subgame does only concern Alice’s preferences: if Alice is self-interested, she will aim at reaching the $(A, A)$ outcome, whereas if she considers the social welfare of the group, she might wish to reach the $(B, B)$ outcome. On the other hand, Bob’s preference orderings in this subgame are the same no matter whether he wishes to maximize his self-interest or the social welfare.
Similarly, considering a Hi-Lo-like subgame would not match our current study as it does not offer any dilemma between satisfying self-interest and maximizing the social welfare: in fact, in the Hi-Lo matching game (see Figure 4.2(b)), both players always obtain the same payoff no matter the outcome, which explains the high rate of coordination on the most profitable outcome for both players, independently of whether there exists a tie between them.

One may also notice the similarity of our Entrance game with the Dalek game presented in Binmore and Samuelson [1999] (a corresponding subgame is depicted in Figure 4.2(c)). The main difference is that one solution of the coordination subgame ensures perfect equity in the Dalek game. Indeed, as in our case, the Dalek game also introduces some dilemma between maximizing one’s self-interest and playing the fairest outcome. However, unlike in our Entrance game, it does not introduce any dilemma between satisfying self-interest and maximizing the social welfare.
welfare (i.e., the combined payoffs of every player). Although this game would be interesting to investigate, focusing on it may also make it more difficult to observe the actual effects of social ties on behavior: as a consequence, the absence of any clear incentive to play the fairest solution in the Dalek game may eventually lead to a higher rate of miscoordination, independently of the presence of such ties. On the other hand, the signal of perfect equity in the Dalek game may also appear so strong that it could reinforce the stability of coordinating on the corresponding solution (i.e., \((B, B)\)), even when no ties are involved.

### 4.4 Equilibrium predictions with self-regarding players

Through this section, we wish to provide a full theoretical analysis of the above Entrance game that is exclusively based on traditional game theory (i.e., assuming agents are self-interested maximizers). Through the rest of this dissertation, we note \(A\) and \(B\) to respectively stand for sub-strategies “\(A\) if \(In\)” and “\(B\) if \(In\)” in the context of the Entrance game.

#### 4.4.1 Nash equilibria

First consider the coordination subgame alone (i.e., the second stage of the full Entrance game). In such a game, both \((A, A)\) and \((B, B)\) are the only pure Nash equilibria, which also appear to be the only Pareto optimal solutions. There also exists a Nash equilibrium in mixed strategy, which consists in playing \(A\) with probability \(7/8\) for Alice, and playing \(B\) with probability \(7/10\) for Bob (in this case, the respective expected payoffs are 10.5 for Alice, and 4.375 for Bob).

The main features of this game lie on defining the role played by the group’s preferences in the players’ behavior. As in the BoS game, being self-interested is not sufficient to guarantee any coordination success: every action is indeed compatible with some common belief in the players’ rationality. However, in the coordination game, one can notice the existence of a focal point for the group that is not present in the classical BoS game: out of the two Nash equilibria, the outcome \((B, B)\) is always better for the group. In fact, no matter whether one considers the sum, the average, the difference, or the minimum value among the individual payoffs as a measure of the group’s utility, this unique outcome always outperforms every other solution. In fact, the asymmetry in the players’ payoffs creates some incentives for them to favor the group as a whole, which can also allow them to eventually maximize their self-interest (any coordination is always better than miscoordination). Both players may then consider this solution as a focal point that can be used to reach coordination. However, one should note that,
as the corresponding solution \((B, B)\) favours Bob more than it favors Alice (what is best for the group is also best for him), the players may still choose to deviate from it. Is Alice likely to detect and follow this focal point \((B, B)\), which clearly conflicts with her best outcome (i.e., \((A, A)\))? What can weaken/strengthen the revealing of this focal point to the players? These are the questions we wish to answer through the experimental study.

Let us now consider the full Entrance game, which consists of the previous coordination game extended with some outside option (at the first stage of the game). The corresponding game in normal form is represented in Figure 4.3.

\[
\begin{array}{ccc}
 & A & B \\
(In, A) & (35,5) & (0,0) \\
(In, B) & (0,0) & (15,35) \\
(Out, A) & (20,10) & (20,10) \\
(Out, B) & (20,10) & (20,10) \\
\end{array}
\]

Figure 4.3: Social Tie game in normal form

This game contains three Nash equilibria in pure strategies, which are the following:

\((In, A), (Out, A; B), (Out, B; B)\)

These equilibria should simply be understood as follows: as long as Bob does play \(B\), then \(Out\) remains the best option for Alice (no matter what Alice would have chosen between \(A\) and \(B\)). In any other cases, strategy \((In, A)\) becomes the only rational move for Alice.

Moreover, the Entrance game also has a Nash equilibrium in mixed strategy, which consists in Alice always playing \(Out\) (i.e., with probability 1) and Bob playing \(B\) with probability \(3/7\). This solution however has to be distinguished from another Nash equilibrium in behavioral strategy, which consists in Alice always playing \(Out\) first (i.e., with probability 1) and playing \(B\) with probability \(1/8\) in the subgame while Bob then plays \(B\) with probability \(7/10\) (Note that this corresponds to the Nash equilibrium in mixed strategy in the subgame). However, one should note that all Nash equilibria in mixed or behavioral strategies are simply irrelevant to the Entrance game: if Alice is willing to randomize in the subgame or believes that Bob will, then she is always better off by playing \(Out\) in the first
place. In this case, the respective expected payoffs are 20 for Alice, and 10 for Bob.

4.4.2 Subgame perfect Nash equilibria

The subgame perfect equilibria, which can be computed through the backward induction method, represent a restriction on the previous set of Nash equilibria. In fact, this solution concept allows to rule out incredible solutions that may be predicted as Nash equilibria. In our game, \((Out, A; B)\) represents such a solution. Indeed, although the prediction to play \(Out\) is perfectly rational for Alice, it here relies on the fact that she would not be rational if she had played \(In\) in the first place: given that Bob plays \(B\) in the coordination subgame, Alice’s only rational move would be to play \(B\) instead of \(A\) (which corresponds to a Nash equilibrium in the subgame).

Moreover, one should note that the backward induction principle also discards the Nash equilibrium in mixed strategies from the previous section. On the other hand, the Nash equilibrium in behavioral strategies remains.

As a consequence, the set of all subgame perfect Nash equilibria in pure strategies reduces to the following:

\[(In, A; A), (Out, B; B)\]

4.4.3 Forward induction

Similarly the forward induction principle restricts the previous set of subgame perfect Nash equilibria to those solutions, which resist the iteration of weak dominance. In the context of our Entrance game (see Figure 4.3), this leads to the following solution: first Alice’s strategy \((In, B)\) is weakly (and strictly) dominated by any strategy involving \(Out\). Then Bob’s strategy \(B\) becomes weakly dominated by \(A\). Thus Alice’s strategies \((Out, A)\) and \((Out, B)\) are both weakly (and strictly) dominated by \((In, A; A)\). Therefore, the unique forward induction solution, which resist iterated weak dominance, is as follows:

\[(In, A; A)\]

Indeed, this solution can be interpreted as follows: while playing \(In\), Alice signals Bob that she intends to play \(A\) (if she intended to play \(B\), she would have played \(Out\) in the first place). Therefore Bob’s unique rational move is to play \(A\). However, while this interpretation justifies the existence of the above solution, it does not explain why the other backward induction solution is not rational. To continue the argument, let us then consider the solution \((Out, B; B)\), which can be interpreted as follows: Alice plays \(Out\) because she expects Bob to play \(B\) in case
she had played *In*. This chain of reasoning is clearly erroneous because Alice’s conditional expectation does not match what she would really expect if she had *actually* chosen to perform *In*. Indeed, as shown before, if Alice performs *In*, Bob’s unique rational move is to play *A*, thus no matter what Alice does during the first stage, she cannot expect anything else than Bob playing *A*. Consequently, her unique rational move is to play (*In*, *A*), and Bob’s best response is to play *A*.

The interesting characteristics that this analysis brings about is that the validity of this forward induction argument is independent of Bob’s preferences. This therefore suggests that such a game introduces some “first mover” advantage, assuming that it is common knowledge among them that they both are self interested agents. One should note that no equilibrium in mixed or behavioral strategies resists this principle.

Many studies in the experimental economic literature have provided support to this forward induction argument, see, e.g., Balkenborg [1994]; Brandts and Holt [1989, 1995]; Cachon and Camerer [1996]; Cooper et al. [1992, 1993]; Shahriar [2009]; Van Huyck John et al. [1993].

One of the first papers in this direction is Brandts and Holt [1989]. Cooper et al. [1992] investigates a coordination game with two Pareto-ranked equilibria and report that a payoff-relevant outside option changes play in the direction predicted by forward induction. Van Huyck John et al. [1993] reports the success of forward induction in a setup in which the right to participate in a coordination game is auctioned off prior to play. Cachon and Camerer [1996] investigates a setup in which subjects may pay a fee to participate in a coordination game with Pareto-ranked equilibria. They report that play is consistent with forward induction.

While many experiments support the fact that people’s strategic behavior relies on the forward induction argument (see, e.g., Cooper et al. [1992, 1993]; Van Huyck John et al. [1993], there is also contrary evidence. In Cooper et al. [1993], Cooper et al. obtain the forward induction solution when it coincides with a dominance argument but the same outcome is predicted when forward induction makes no prediction. Brandts and Holt [1995] also shows that the forward induction is a good prediction only if it coincides with a simple dominance argument. In Brandts et al. [2003], the author find evidence against forward induction in an industrial organization game.

Other work have shown that the temporal factor of the game is relevant to forward induction reasoning. In Cooper et al. [1993] and Huck and Muller [2005], the forward induction solution predicts well subjects’ behavior in an experimental game in extensive form, but does poorly when subjects are presented with the normal form game. A similar problem seems to arise in Caminati et al. [2006] who analyse games similar to ours but who work essentially with the normal form.

However, all these works consider games that are slightly different from the
interactive strategic situation on which we focus in this chapter. One may then wonder whether the asymmetry introduced in our Entrance game does alter the game theoretic prediction.

4.5 Equilibrium predictions under models of social preferences

In this section, we reinterpret our Entrance game through the use of well-known economic theories of social preferences and analyze players’ equilibrium behavior under these theories. In fact, these models allow one to consider not only the self-interested motivations of the players, but also their social motivations, which may then be particularly important in the context of social ties. In other words, a player’s utility is not characterized by his own material payoff only, but also those of the other players. We choose to focus on the concepts of inequity aversion and fairness, which seem to be the most relevant to our Entrance game. Other models of intentions-based fairness and reciprocity (see, e.g., Rabin [1993a]) do not appear to be suitable to such a coordination game. Apart from the problem of multiple equilibria in beliefs that characterizes such belief-dependent approaches, it would be difficult to unambiguously define what is “kind” and what is “unkind” in the players’ strategy set, by using only first and second order beliefs.

4.5.1 Theory of inequity aversion

In the models proposed in Fehr and Schmidt [1999] and Bolton and Ockenfels [2000], players are assumed to be intrinsically motivated to distribute payoffs in an equitable way: a player dislikes being either better off or worse off than another player. In other terms, utilities are calculated in such a way that equitable allocations of payoffs are preferred.

Formally, consider two players \(i\) and \(j\) and let \(x = \{x_i, x_j\}\) denote the vector of monetary payoffs. According to Fehr & Schmidt’s inequity aversion model, the utility function of player \(i\) is given by:

\[
U_{IA}^i(x) = x_i - \alpha_i \cdot \max\{x_j - x_i, 0\} - \beta_i \cdot \max\{x_i - x_j, 0\}
\]

where it is assumed that \(i \neq j\), \(\beta_i \leq \alpha_i\) and \(0 \leq \beta_i < 1\).

The two parameters can be interpreted as follows: \(\alpha_i\) parametrizes the distaste of person \(i\) for disadvantageous inequality while \(\beta_i\) parametrizes the distaste of person \(i\) for advantageous inequality. One should note that setting these parameters to zero defines some purely self-interested agent. The constraints imposed on the parameters are meant to ensure that players cannot distaste advantageous inequality more than disadvantageous inequality in order to be realistic.
Clearly, applying such a model to our current Entrance game can literally transform its whole structure, depending on the values assigned to parameters $\alpha_i$ and $\beta_i$. Let us then perform a game theoretic analysis that involves such inequity aversion parameters.

The main observation that can be made is about the effects of Alice’s preference ordering on her behavior. In fact, assuming that $\beta_a \leq \alpha_a$, then Alice will never play the strategy $(I, B)$, no matter how inequity averse she is:

- if $\beta_a < 3/4$ and $\alpha_b < 1/6$, then Alice and Bob’s optimal behavior remains as if they were self-interested (i.e., the forward induction argument still holds). Thus Alice’s unique rational strategy is to play $(I, A)$ while Bob will rationally play $A$.

- if $\beta_a < 3/4$ and $\alpha_b \geq 1/6$, then Alice is always better off by playing $(O, \cdot)$: the coordination subgame yields a unique Nash equilibrium (i.e. $(B, B)$), which is strictly dominated by strategy $(O, \cdot)$.

- if $\beta_a \geq 3/4$, then Alice is always better off by playing $(O, \cdot)$: for any $\alpha_a \geq \beta_a$, any outcome from the coordination subgame is strictly dominated by playing $(O, \cdot)$ (see Figure 4.4 for an example).

Figure 4.4: Transformed Entrance game with extremely inequity averse players ($\alpha_a = \beta_a = \alpha_b = \beta_b = 1$)

The main result of this analysis is that the value of $\alpha_a$ and $\beta_b$ are irrelevant to defining Alice and Bob’s optimal behavior. In other words, only Alice’s distaste
about advantageous inequality can affect her preference ordering in the Entrance game. Similarly, only Bob’s distaste about disadvantageous inequality can affect his decision. One should also note that inequity aversion does not keep the “first mover” advantage mentioned in the previous section: Alice’s first move does signal Bob not only about her low level of inequity aversion, but also about her expectation of Bob’s low level of inequity aversion. That means that if she plays \( \text{In} \), then the resulting outcome is entirely depending on Bob’s level of inequity aversion (either \((\text{In}, \text{A}; \text{A})\) or \((\text{In}, \text{A}; \text{B})\) will be played).

The sets of Nash Equilibria (NE) and of Subgame Perfect Nash Equilibria (SPE), in the context of the Entrance game played with inequity aversion, are summarized in Table 4.1 (note that forward induction is irrelevant in this case because the SPE is unique for every vector \((\alpha_i, \beta_i)\), with \(i \in \{a, b\}\) and \(\alpha_i, \beta_i \in [0, 1]\)).

<table>
<thead>
<tr>
<th>NE</th>
<th>SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\text{Out}, \text{A}; \text{A}))</td>
<td>((\text{Out}, \text{A}; \text{A})) if (\alpha_b &lt; 1/6)</td>
</tr>
<tr>
<td>((\text{Out}, \text{B}; \text{B}))</td>
<td>((\text{Out}, \text{B}; \text{B})) if (\beta_a &lt; 3/4)</td>
</tr>
<tr>
<td>((\text{Out}, \text{A}; \text{B}))</td>
<td>((\text{Out}, \text{A}; \text{B})) if (\alpha_b \geq 1/6) and (\beta_a \geq 3/4)</td>
</tr>
<tr>
<td>((\text{Out}, \text{B}; \text{A}))</td>
<td>((\text{Out}, \text{B}; \text{A}))</td>
</tr>
</tbody>
</table>

Table 4.1: Equilibrium solution concepts for inequity averse agent(s) \((\beta_a \geq 3/4\) or \(\alpha_b \geq 1/6\))

4.5.2 Theory of fairness

Let us now consider another type of social preferences model, that in turn relies on the notion of fairness. In Charness and Rabin [2002], Charness & Rabin propose a specific form of social preference they call quasi-maximin preferences. In their model, group payoff is computed by means of a social welfare function which is a weighted combination of Rawls’ maximin and of the utilitarian welfare function (i.e., summation of individual payoffs) (see [Charness and Rabin, 2002, p. 851]).

Formally, consider two players \(i\) and \(j\) and let \(x = \{x_i, x_j\}\) denote the vector of monetary payoffs. According to Charness & Rabin’s fairness model, the utility function of player \(i\) is given by:

\[
U^F_i(x) = (1 - \lambda) \cdot x_i + \lambda \cdot [\delta \cdot \min\{x_i, x_j\} + (1 - \delta) \cdot (x_i + x_j)]
\]

where \(\delta, \lambda \in [0, 1]\). Moreover, the two parameters can be interpreted as follows: \(\delta\) measures the degree of concern for helping the worst-off person versus maximizing the total social surplus. Setting \(\delta = 1\) corresponds to a pure “maximin” (or “Rawlsian” criterion), while setting \(\delta = 0\) corresponds to total-surplus
maximization. Parameter $\lambda$ measures how much player $i$ cares about pursuing the social welfare versus his own self-interest. Setting $\lambda = 1$ corresponds to purely “disinterested” preferences, with player $i$ caring no more (or less) about her own payoffs than others’, while setting $\lambda = 0$ corresponds to pure self-interest.

As for the previous model, the parameters $\delta$ and $\lambda$ can considerably change the structure of the Entrance game, which is why we propose a new game theoretic analysis involving such fair agents.

The first observation is that while fairness may slightly alter Bob’s preferences, the $(In, B; B)$ outcome always remains the best option: the only difference with the classical model is that he may come to prefer the $(In, A; A)$ outcome to the $(Out, \cdot)$ outcome when $\delta < 2/3$ and $\lambda > 1/3$.

Similarly, Alice’s preferences also get affected by such notion of fairness. The main result is that a new forward induction solution may emerge through such a social preferences model. In particular:

- if $\lambda < 1/2$, then Alice may still play the forward induction solution strategy as predicted by traditional game theory (i.e., $(In, A)$), depending on the value of $\delta$.

- if $1/2 \leq \lambda \leq 3/4$, then no prediction can be made without considering probabilistic beliefs: both Nash equilibria in pure strategies in the subgame are always at least as good for Alice as playing $(Out, \cdot)$.

- if $\lambda > 3/4$ and $\delta > 2/3$, then Alice may play a forward induction solution strategy (i.e., $(In, B)$) that mainly relies on her other regarding preferences (see Table 4.2): solution $(In, B; B)$ indeed becomes preferred to playing $(Out, \cdot)$, which is preferred to solution $(In, A; A)$ (see Figure 4.5 for an example).

Moreover, one should note that, as for the original version of the game (see section 4.4), the $Out$ option for Alice always dominates the Nash equilibrium in mixed strategies from the coordination subgame, no matter what the values of $\lambda$ and $\delta$ are.

The above analysis suggests that the Entrance game may in fact contain two distinct focal points for the players, which can be identified by the two possible forward induction solutions. Therefore, one can state that the current Entrance game yields a unique social-welfare equilibrium\(^1\) if and only if players have either some strong self-interested preferences ($\lambda << 1/5$) or some strong other-regarding preferences ($\lambda >> 3/4$ and $\delta >> 2/3$). In the latter case, one should note that

\(^1\)The social welfare equilibrium introduced by Charness & Rabin ([Charness and Rabin, 2002, p. 852]) corresponds to a Nash equilibrium for some given values of $\delta$ and $\lambda$. 

88
Figure 4.5: Transformed Entrance game for extremely fair agents ($\lambda = \delta = 1$)

the players’ sensibility to the maximin principle needs to “dominate” that of the utilitarian welfare function.

The sets of Nash Equilibria (NE), of Subgame Perfect Nash Equilibria (SPE), and of Forward Induction solutions (FI), in the context of the Entrance game played by fair agents, are shown in Table 4.2.

### 4.6 Equilibrium predictions under our model of social ties

Similarly to the theories of social preferences considered in the previous section, our main claim is that the strength of the social tie existing between two players has some important effects on their preferences (and consequently on their expected behavior), as suggested in Section 4.2. However, as we believe that the type of payoff transformation used in our model is more appropriate to the context of

<table>
<thead>
<tr>
<th>NE</th>
<th>SPE</th>
<th>FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(In, B; B)$</td>
<td>$(In, B; B)$</td>
<td>$(In, B; B)$</td>
</tr>
<tr>
<td>$(Out, A; A)$</td>
<td>$(Out, A; A)$</td>
<td></td>
</tr>
<tr>
<td>$(Out, B; A)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Equilibrium solution concepts for fair agents ($\lambda \gg 3/4$ and $\delta \gg 2/3$)
social ties, we illustrate through this section how it can disagree with the previous concepts from Section 4.4 and Section 4.5 in terms of equilibrium predictions. More specifically, let us apply our model of social ties to the above Entrance game.

Given the subjective social ties values $k_{ab}$ and $k_{ba}$, and the constant values $k_1, k_2, k_3 \in [0, 1]$ such that $k_1 \leq k_2 \leq k_3$, we have:

- if $k_{ab} > k_3$, then the unique rational play for Alice and Bob is to coordinate on $(In, B; B)$, independently of $k_{ba}$.
- if $k_{ab} > k_1$ and $k_{ba} > k_2$, then, as in the previous case, the unique Nash equilibrium is again for both players to coordinate on the $(In, B; B)$ outcome (see Figure 4.6 for an example).
- if $k_{ab} \leq k_1$ and $k_{ba} > k_2$, then Alice should play $(Out, \cdot)$, in response to Bob playing $B$ in the subgame.
- if $k_{ab} \leq k_1$ and $k_{ba} \leq k_2$, then Alice and Bob should follow forward induction reasoning and play $(In, A; A)$. In this case, as the strategic structure of the game remains as in its original version, the game-theoretic analysis from Section 4.4 still applies.
- if $k_1 < k_{ab} \leq k_3$ and $k_{ba} \leq k_2$, then both players are unable to coordinate on a unique Nash equilibrium outcome in the subgame: both $(A, A)$ and $(B, B)$ are Nash equilibria. As a result of such indecision in the subgame, Alice’s optimal strategy is $(Out, \cdot)$.

\[
\begin{array}{|c|c|}
\hline
& A & B \\
\hline
(In, A) & (5,10) & (5,15) \\
(In, B) & (15,10) & (15,15) \\
(Out, A) & (10,10) & (10,15) \\
(Out, B) & (10,10) & (10,15) \\
\hline
\end{array}
\]

Figure 4.6: Transformed Entrance game for socially tied agents ($U = U_m, k_{ab} = k_{ba} = 1$)

In the above analysis, the constant values for $k_1$, $k_2$, and $k_3$ depend on whether the players follow the maximin function (i.e., $U = U_m$) or the utilitarianism principle (i.e., $U = U_s$). Table 4.3 provides the corresponding constant values for each
of these types. Moreover, this analysis considers the most general case where ties may be unilateral (i.e., where \( k_{ab} \neq k_{ba} \)): our current model therefore allows to state that, for instance, Alice feels close to Bob while Bob does not feel close to Alice.

One should note from this interpretation that Alice will always play her strategy \((In, B)\) whenever \(k_{ab} > k_3\), and similarly, Bob will always play his strategy \(B\) whenever \(k_{ba} > k_2\). Such an observation, combined with the fact that \(k_3 > k_2\), indicates that Alice’s decision is more restrictive than Bob’s: if \(k_2 < k_{ab}, k_{ba} < k_3\), then Alice needs to take Bob’s decision into account in order to make her decision, whereas Bob will play \(B\) independently of Alice’s action. As a consequence, introducing social ties in the context of this game may allow Bob’s threat of playing \(B\) to become more credible to Alice’s eye. In other words, social ties may simply turn Alice’s first mover advantage (as suggested in Huck and Muller [2005]) in Bob’s favour.

One should also note the distinction between using utilitarianism or egalitarianism in the tie utility function. As shown in Table 4.3, utilitarianism allows to coordinate more easily on the \((In, B; B)\) outcome than egalitarianism (i.e., \(k_1, k_2,\) and \(k_3\) have lower values when \(U = U_s\)). On the other hand, players following egalitarianism are expected to coordinate in the subgame more often than in the case of utilitarianism (i.e., when \(U = U_m\), Alice should play \((Out, \cdot)\) if and only if \(k_{ab} = k_1 = k_2 = 1/2\)).

However, one may state that the formulation of social ties from Section 4.2 is too general with respect to the concept presented in Section 4.1. In fact, according to Statement 4.1.0.1, we assume that a social tie is restricted to be bilateral, which may not always be the case in the above analysis. In order to match this criterion, one then needs to add the following constraint:

\[
k_{ab} = k_{ba}
\]

In this case, assuming \(k = k_{ab} = k_{ba}\), coordination on the \((In, B; B)\) outcome is reached only when \(k > k_2\). Similarly, coordination on the \((In, A; A)\) outcome is reached only when \(k < k_1\). On the other hand, Alice will play \((Out, \cdot)\) whenever \(k_1 \leq k \leq k_2\), as miscoordination would be expected in the subgame. As this

<table>
<thead>
<tr>
<th>Constant</th>
<th>Egalitarianism</th>
<th>Utilitarianism</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k_i)</td>
<td>(U = U_m)</td>
<td>(U = U_s)</td>
</tr>
<tr>
<td>(k_1)</td>
<td>1/2</td>
<td>1/5</td>
</tr>
<tr>
<td>(k_2)</td>
<td>1/2</td>
<td>1/3</td>
</tr>
<tr>
<td>(k_3)</td>
<td>4/5</td>
<td>7/9</td>
</tr>
</tbody>
</table>

Table 4.3: Constant tie values for each type of player
constraint simplifies decision making for both players in the Entrance game, one should note that it also removes any opportunity for Bob to exploit Alice: whenever $k_2 < k_{ab}$, Alice will always play $(In, B)$ independently of Bob’s decision.

More specifically, about determining Alice and Bob’s dinner plan, this analysis leads to the following interpretation. On the one hand, if the existing social relationship is sufficiently weak between Alice and Bob, then Bob should be influenced by Alice’s intention of going to her favourite Japanese restaurant and, consequently, he should choose $A$ in order to maximize her own payoff. On the other hand, in the presence of a sufficiently strong social tie between them, Alice should be influenced by Bob’s intention of doing what is best for the group and, consequently, she should choose $B$ in order to maximize the group payoff.

Furthermore, in the particular case of an intermediate measure of social ties between Alice and Bob, each individual may then become uncertain about the other’s choice of either acting fully rationally and go to Alice’s favourite restaurant, or doing what is best for the group and go to Bob’s favourite restaurant. In response to such a high risk of meeting at the wrong restaurant, Alice should then reaffirm her intention to stay eat at her place with Bob.

However, one should note that the above constraint on both players’ subjective social ties (i.e., $k_{ab} = k_{ba}$) suggests the existence of a common scale for measuring such ties. In fact, in order for Alice to determine the actual social tie level between Bob and herself, then she must first make sure that her notions of the weakest and highest possible ties are the same as Bob’s (e.g., does “being best friends” have the same meaning for both of them?). For example, it is fair to say that everybody does not share the same level of tie with a complete stranger (some are genuinely more cooperative than others). One may then argue that the non-easy task of normalizing one’s social ties scale is already part of what defines the social tie itself: indeed the higher the social tie value, the more the concerned individuals are likely to share the same social ties scale (e.g., if I feel that we are best friends, but I ignore what “being best friends” means to you, then I will not risk being exploited by you, and my social tie with you will consequently remain weak). As a result, this analysis follow our definition of social ties from Section 4.1 as it suggests that social ties are intrinsic psychological factors that are influenced by the agents’ epistemic states (the more I know about you, the more my measure of our social tie is reliable).

### 4.7 Players as team-directed reasoners

Our proposed model of social ties from Section 4.2 appears to share some common properties with another well-known concept that also relies on group identification: team reasoning. We therefore provide, in this section, a detailed analysis of the
Entrance game through the various theories of team reasoning as a means to illustrate the common characteristics as well as the differences that exist with our model of social ties.

One should note that, in the context of the Entrance game, considering collective utility functions (see, e.g., the classical utilitarianism and the *maximin* principle from Section 4.2) from the players' individual viewpoint can lead to a transformed game similar to that depicted in Figure 4.5 from Section 4.5.2. In this case, the resulting subgame in Figure 4.5 has similar properties as the well known Hi-Lo matching game: as both players have the same preferences over outcomes, they indeed benefit if and only if they coordinate with each other in the subgame. However, their subsequent payoffs depend on which action they do coordinate on. The interesting property of this transformed subgame is that it introduces a dilemma that even economic theory cannot solve. However, while game theory is indeed unable to predict any particular outcome (i.e., both coordinated outcomes of the subgame are Nash solutions), it is shown in Bacharach [2006] that people would tend to coordinate on the action that leads to the most rewarding outcome for both, i.e., \((B, B)\).

In order to interpret such intuitive behavior, some theorists have proposed to incorporate new modes of reasoning into game theory. For instance, starting from the work of Gilbert (Gilbert [1989]) and Reagan (Regan [1980]), some economists and logicians (e.g., Lorini [2011]) have studied team reasoning as an alternative to the best-response reasoning assumed in traditional game theory (Bacharach [1999]; Colman et al. [2008]; Sugden [2000, 2003]). Team-directed reasoning is the kind of reasoning that people use when they perceive themselves as acting as members of a group or team (Sugden [2000]). That is, when an agent \(i\) engages in team reasoning, he identifies himself as a member of a group of agents \(S\) and conceives \(S\) as a unit of agency acting as a single entity in pursuit of some collective objective. A team reasoning player acts for the interest of his group by identifying a strategy profile that maximizes the collective payoff of the group, and then, if the maximizing strategy profile is unique, by choosing the action that forms a component of this strategy profile.

Furthermore, as suggested in Hakli et al. [2010], the concept of team reasoning also refers to Tuomela’s *I-mode / we-mode* distinction from Tuomela [2010]. According to Tuomela, the *I-mode* consists in reasoning as a private person according to two possible principles: an agent reasoning in *plain I-mode* will seek to satisfy self-interest as suggested by classical economic theory (this corresponds to the type of reasoning underlying the analysis from Section 4.4). On the other hand, an agent reasoning in *pro-group I-mode* is concerned with promoting the group’s interests, and as a result will make a decision with the individual intention to maximize the group utility. As an example of such a benefactor behavior (as
called by Bacharach in Bacharach [1999]), one may consider the existing theories of social preferences presented in Section 4.5 (e.g., the concept of social welfare equilibrium presented in Charness and Rabin [2002] illustrates this type of thinking). It is clear however that even such pro-group I-mode thinking, which relies on preference transformation, fails to predict some very intuitive behavior such as in the well known Hi-Lo matching game, as shown in Bacharach [2006]; Colman et al. [2008] (see Figure 4.5 for a similar example). The alternative concept of we-mode reasoning instead relies on what Bacharach calls agency transformation, which consists in conceiving the situation not as a decision making problem for individual agents (cf. the I-mode), but as a decision making problem for the group conceived as an agent.

Let us now perform a detailed analysis of team reasoning applied to the Entrance game according to both Sugden and Bacharach’s different theories\(^1\). Figure 4.7 illustrates a representation of the Entrance game from the group’s viewpoint when considering the maximin principle as the group utility function. In this case, the transformed Entrance game considers a unique player, which corresponds to the group \{Alice, Bob\}. We indicate with \(s_a,s_b\) any group’s strategy. One therefore notes that the best strategy for the group is to always play \((\text{In},B;B)\).

First, according to Sugden’s theory (Gold and Sugden [2007]; Sugden [2003]), a simple epistemic interpretation of team reasoning (from Alice’s viewpoint) in the

\(^1\)A more detailed general comparison of Sugden and Bacharach’s theories of team reasoning can be found in Gold [2012].
current Entrance game can be the following (for $U \in \{U_s, U_m\}$):

**Statement 4.7.0.1** If Alice believes that:

- She is a member of the group $\{Alice, Bob\}$.
- It is common knowledge among Alice and Bob that both identify with $\{Alice, Bob\}$.
- It is common knowledge among Alice and Bob that both want the value of $U$ to be maximized.
- It is common knowledge among Alice and Bob that $(In, B; B)$ uniquely maximizes $U$.

Then she should choose her strategy $(In, B)$.

Following Statement 4.7.0.1, it is then clear that if Alice shares her beliefs with Bob, then the resulting outcome will be $(In, B; B)$ (i.e., Bob will similarly choose the corresponding option $B$). According to Sugden (Sugden [2003]), a player has reason to act as a team member and to choose the action that forms a component of the strategy profile maximizing collective payoff, conditional on assurance that the other players also act as team members. That is, to act as a member of a team, one must be confident that the other players act as members too. More fundamentally, “[...] team reasoning does not generate reasons for choice unless each member of a team has reason to believe that there is common reason to believe that each member of the team endorses and acts on team reasoning [...] This is a condition of assurance” ([Sugden, 2003, p. 176-177]). In other words, the main characteristics of Sugden’s theory is that team reasoning relies on strong epistemic foundations and is very restrictive in that matter: an agent will not take the risk to team-reason and be “suckered” by other agents who do not.

Let us now consider Bacharach’s theory of team reasoning as an alternative to Sugden’s previous interpretation, which, as shown in Hakli et al. [2010], yields the same action recommendations as Tuomela’s *we*-mode reasoning in any game-theoretic situation. In Bacharach [1999], Bacharach introduces the concept of unreliable team interaction, which corresponds to a game structure in which there is a probability that a given player identifies with a team and chooses the action which maximizes the team benefit (i.e., the player plays in the *we*-mode), and another probability that the player is a self-interested agent who tries to maximize his own benefit (i.e., the player plays in the *I*-mode). In this sense, the interaction is “unreliable” because there is no certainty that a player will reason and act as a team member. A team member will then act according to his expected utility based on what others will do (including players who do not team-reason). In such an unreliable team interaction, Bacharach also introduces the notion of a team
Table 4.4: Protocol Equilibria (with $X_a \in \{Out, InA, InB\}$ and $X_b \in \{A, B\}$)

<table>
<thead>
<tr>
<th>Probability we-mode</th>
<th>Protocol Equilibria (for $U \in {U_m, U_s}$)</th>
</tr>
</thead>
</table>
| $\omega = 0$        | $(InA, A, X_a X_b)$  
|                     | $(Out, B, X_a X_b)$  |
| $0 < \omega \leq 3/4$ | $(InA, A, InA A)$ if $U = U_s$  
|                     | $(InA, A, Out A)$ if $U = U_m$  
|                     | $(Out, B, InB B)$  |
| $3/4 < \omega \leq 8/9$ | $(InA, A, InA A)$ if $U = U_s$  
|                     | $(Out, B, InB B)$  |
| $8/9 < \omega < 1$ | $(Out, B, InB B)$  |
| $\omega = 1$        | $(X_a, X_b, InB B)$ |

protocol, which consists in specifying a strategy for every player when identifying with each team (or in each mode). As an example, the protocol $(A, B, BA)$ specifies that Alice and Bob will respectively play $A$ and $B$ if in $I$-mode, and will respectively play $B$ and $A$ if in $we$-mode. Through this concept of team protocol, Bacharach differentiates an agent’s behavior depending on whether he identifies with the group or not. The players then reach a protocol equilibrium if and only if, given the probability $\omega$ that each player reasons in $we$-mode, neither players nor the group can increase its expected utility by individually deviating from it. In other words, such a protocol equilibrium may simply be interpreted as a Nash equilibrium in the extended game where the group $\{Alice, Bob\}$ becomes an extra player. Such a property illustrates a major difference between Bacharach’s theory and our model of social ties. In fact, while an unreliable team interaction requires to consider additional players in the game in order to perform strategic reasoning, each of which corresponding to a combination of individual players (cf. the concept of a team protocol), our model of ties instead only leads to a modification of each individual’s utility, leaving the game structure in its original version (i.e., the sets of players and strategies remain unchanged).

Table 4.4 describes the sets of Protocol Equilibria (PE) for each probability value $\omega$ that each player reasons in $we$-mode in the Entrance game.

One can see from Table 4.4 that, when the agents reason in $I$-mode (i.e., $w = 0$), the set of protocol equilibria matches that of Nash equilibria. Conversely, if both agents play in $we$-mode (i.e., $w = 1$), then the only equilibrium is to play the solution $(In, B; B)$. Moreover, one can note that the maximin principle requires a lower probability of $we$-mode reasoning ($w > 3/4$) than the utilitarianism principle ($w > 8/9$) in order to converge to this unique solution. In other words, the utilitarianism principle requires a stronger identification with the same group in
order to achieve coordination. One should note that team reasoning agrees with our model of social ties regarding the behavioral predictions in the context of the Entrance game.

However, the main limitation of Bacharach’s theory is that it does not clarify what the probabilistic distribution \( \omega \) stands for in the definition of an unreliable team interaction structure. In fact, while such probabilities may depend on some intrinsic features of the game such as the payoff structure, they may also reasonably be determined by some pre-existent social relationships between the players: two strongly (resp. weakly) tied individuals may indeed each have a high probability of being in we-mode (resp. I-mode) in situations like the above Entrance game.

To further the analysis, let us note that the concept of unreliable team interaction can be seen as a special type of incomplete information games\(^1\) where the only uncertainty one can have is regarding the level to which other players identify with different groups (e.g., agent \( i \) may identify with the group \( \{i, j\} \) with probability \( \omega \) or with the group \( \{i\} \) with probability \( 1 - \omega \)). In other words, this theory relies on the assumption that every agent identifies with a unique team at a given time, which is a strong assumption. This observation therefore raises the issue of the endogenous determination of the mode of reasoning (i.e., I-mode/we-mode).

In fact, a fundamental point in Bacharach’s original theory, in contrast to Sugden’s theory, is that the determination of mode of reasoning is a psychological matter, prior to any rational choice, and such a process is based on frames. A frame, as first introduced in Bacharach and Bernasconi [1997] through the Variable Frame Theory (VFT), can be defined as a set of concepts that an agent uses when thinking about a decision problem: a person may then start to we/I-reason only if he has ‘we’/‘I’ concepts in his frame, which leads him to answer the corresponding question “What shall we/I do?”. While Bacharach’s theory assumes an agent can only use one frame at once (i.e., an agent cannot reason in I-mode and we-mode at the same time), it is suggested in Smerilli [2010] that some vacillation between different frames may actually occur in one’s mind when facing a decision problem. The corresponding model indeed defines the probability \( \omega \) as a function of the probability of vacillating from we-mode to I-mode and the probability of vacillating from I-mode to we-mode.

Applying such a model to our current Entrance game leads to the following interpretation: if Alice and Bob start by we-reasoning, there will be a unique we-equilibrium \( (\text{In}, \text{B}; \text{B}) \), which is not a Nash equilibrium. So if Alice starts with we-mode reasoning, she will not be happy with the result and move away from this equilibrium (e.g., by playing \( (\text{Out}, \cdot) \)). If instead Alice starts with the I-mode, both

\[^1\text{Note that we do not refer to the usual Bayesian game as defined by Harsanyi here. It is indeed shown in Hakli et al. [2010] that a Bayesian game generated from Bacharach’s unreliable team interaction structure does not yield the same action recommendation.}\]
individuals shall not be happy (both \((In, A; A)\) and \((Out, B; B)\) are dominated by the previous “we” solution \((In, B; B)\)). In this case there can be a continuous switching or vacillation from one frame to another. As shown in Smerilli [2010], this interpretation is similar to that of the well known prisoner’s dilemma, which therefore suggests that miscoordination should prevail in the Entrance subgame\(^1\). Consequently, one could conjecture that reinforcing social ties between individuals in the Entrance game decreases the probability \(p\) of vacillating between \textit{we-mode} and \textit{I-mode}, while increasing the probability \(q\) of vacillating between \textit{I-mode} and \textit{we-mode}.

\section*{4.8 Why team reasoning cannot express gradual social ties}

Following the previous analysis of team reasoning, one may however wonder whether this interpretation, and more generally Sugden and Bacharach’s theories, are actually adequate to interpret the effects of social ties, as they clearly forbid the possibility that an agent is reasoning in two different modes at the same time. Indeed, in some unpublished work (Bacharach [1997]), Bacharach allows for the existence of some “superordinate” frame where an agent can see the problem from both the ‘I’ and the ‘we’ perspective, even though he states that those perspectives cannot hold simultaneously. In order to better understand Bacharach’s view regarding this matter, one may consider the analogy with Rubin’s vase, which is illustrated in Figure 4.8 (as already suggested in Smerilli [2010]).

\begin{figure}[h]
\includegraphics[width=0.3\textwidth]{rubins_vase}
\caption{Rubin’s vase}
\end{figure}

When looking at the image in Figure 4.8, one can indeed see either a vase (in

\footnote{Experimental results in the Prisoner’s dilemma have shown that the cooperation rate varies between 30-40\% (see e.g., Shafir and Tversky [1992]).}
black color) or two faces (in white color). In this case, one can easily vacillate between perceiving both forms, but one cannot see both concepts simultaneously. In the same fashion, Bacharach reasonably assumes that an individual simultaneously cannot perceive a social situation from the “I” perspective and from the “we” perspective.

However, this interpretation, which is assumed in Sugden and Bacharach’s theories, does clearly not allow to capture the fact that one may identify with a given group up to a certain degree. The need for such a gradual group identification is justified by the various social features that may simultaneously define one’s social identity, as suggested by our basic definition of social ties from Section 4.1. For example, two individuals may consider political orientations (e.g., being a Democrat) and religion (e.g., being a Catholic) as very important social features. In this case, it is reasonable to state that the social tie between them if they share both of these features (e.g., both are Democrat and Catholic) is stronger than if they share only one of those (e.g., both are Democrat, but one is Catholic while the other is Muslim), which is itself assumed to be stronger than sharing none of them (e.g., one is Republican and Catholic while the other is is Democrat and Muslim).

In order to illustrate more formally the differences existing between our model of social ties from Section 4.2 and the concept of team reasoning from Section 4.7, let us consider a simple concrete two player game where player $i$ can choose between three options: $A$, $B$, and $C$. In such a scenario, each player’s payoff is determined uniquely from these options according to Table 4.5 (for simplicity, player $j$ has no control over the outcome). Note that the pair’s payoff function can then follow either utilitarianism (i.e., sum of individual payoffs) or the maximin principle (i.e., minimum of individual payoffs).

<table>
<thead>
<tr>
<th>Player (i)’s option</th>
<th>Payoffs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Player (i)</td>
</tr>
<tr>
<td>$A$</td>
<td>8</td>
</tr>
<tr>
<td>$B$</td>
<td>5</td>
</tr>
<tr>
<td>$C$</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 4.5: Simple dictator game

Applying team reasoning to this particular situation leads to the following predictions: player $i$ will play $A$ if reasoning in I-mode (player $i$ is then self-interested), and player $i$ will play $B$ if reasoning in we-mode (player $i$ then identifies with the group). As a consequence, according to both Bacharach and Sugden’s

---

1The game presented here is overly simplistic as a means to illustrate the above point, which could of course also be found in more classical types of social interactions.
theories of team reasoning (and independently of whether the group utility function follow utilitarianism or the maximin principle), player $i$ will never happen to choose $C$. Indeed, while Sugden’s theory does clearly not allow for such gradual team identification, Bacharach’s theory cannot provide an unreliable team interaction structure with a probability $\omega$ of identifying with the group ($0 \leq \omega \leq 1$) that specifies this outcome to occur.

On the other hand, considering the same game through our model of social ties from Section 4.2 leads to a different interpretation: in this case, player $i$ will select $A$ if both players are extremely close to each other (e.g., $k_{ij} = k_{ji} = 1$), and player $i$ will select $B$ if they instead are perfect strangers (e.g., $k_{ij} = k_{ji} = 0$). However, if both players are neither best friends nor perfect strangers but, say, simple acquaintances (e.g., $k_{ij} = k_{ji} = 0.5$), then player $i$ will choose $C$ (assuming either utilitarianism or the maximin principle as the group utility function), as a compromise between being self-regarding and group-regarding (For a more detailed comparative analysis of this particular game, see Section 6.3.3.1 from Chapter 8, and Appendix C.2). One can therefore observe that the concept of team reasoning has limited expressive power in the context of social ties, as all the theories presented Section 4.7 are unable to make such an intuitive prediction.

In fact, we claim that our model of social ties is more general than team reasoning in the sense that it allows for a gradual measure of group identification, which, we believe, is an important requirement to capture the actual ongoing behavior under the effects of social ties, as suggested by Section 4.1: according to our model, an agent may indeed partially identify with the group while remaining partly self-interested. More generally, our model allows to represent the fact that a person is socially tied with another individual up to a certain degree.

Furthermore, the characteristics of our model of social ties seem even more suitable to realistically represent more flexible and heterogenous multi-player interactions where different coalitions might be formed (depending on the way the agents are socially tied with one another). Such complex games indeed justify the need for a continuous measure of group identification: as social ties are defined as independent values from an individual to another, an agent may then be similarly tied with various individuals at the same time. Consider, for example, a scenario where one faces the dilemma between cooperating with a close friend, and cooperating with a family member. In this case, assuming the friend and the family member do not know each other, it does clearly not make sense for one to reason as a unique team including all individuals. As a result of identifying with either of the sub-groups (i.e., either the one including the friend or the one including the family member), the theories of team reasoning would then predict that one will choose over these two options, even though a more egalitarian solution might exist that would satisfy more equally all players. Under the same condition, our model
of social ties would instead suggest to select the latter solution. We however postpone the study of such complex situations to Chapter 8, which will provide a more detailed comparative analysis between our theory of social ties and Bacharach’s theory of team reasoning.

4.9 Further hypotheses

As previously mentioned, the main goal of our Entrance game is to investigate whether social ties affect social preferences. According to our model of social ties presented in Section 4.2, our first hypothesis is that social ties do not correlate with inequity aversion. Indeed, while our model of ties supports coordination on the \((In, B; B)\) outcome in the presence of a social tie, inequity aversion instead predicts that Alice will play \((Out, -)\), no matter whether she is and/or expects Bob to be inequity averse. However, our model of ties cannot be easily distinguished from the model of fairness presented in Section 4.5.2: both theories may indeed predict the same outcome \((In, B; B)\) can appear to be a unique social welfare equilibrium). In order to discriminate between those models, one may then consider a version of the Entrance game where the outside option is removed (that is, without the possibility for Alice to choose between \(In\) and \(Out\) before the coordination game): this simply corresponds to playing the coordination game alone. Figure 4.9 illustrates it through examples of the corresponding transformed payoff matrix when applying social preferences and the model of social ties to the reduced game.

\[
\begin{array}{c|c|c}
& A & B \\ \hline
A & (5,-25) & (0,0) \\ B & (0,0) & (-5,15) \\
\end{array}
\]
\[
\begin{array}{c|c|c}
& A & B \\ \hline
A & (5,5) & (0,0) \\ B & (0,0) & (15,15) \\
\end{array}
\]
\[
\begin{array}{c|c|c}
& A & B \\ \hline
A & (5,5) & (5,15) \\ B & (15,5) & (15,15) \\
\end{array}
\]

(a) Inequity aversion model (b) Fairness model (c) Social tie model

Figure 4.9: Transformed coordination games without outside option

According to Figure 4.9(a), both Alice and Bob are extremely inequity averse agents (i.e., \(\alpha_a = \alpha_b = \beta_a = \beta_b = 1\)), which leads them to miscoordinate by playing the \((A, B)\) solution. Similarly, considering extremely fair agents (i.e., \(\lambda = \delta = 1\) for Alice and Bob) as in Figure 4.9(b) shows that both players cannot be expected to always coordinate: as the resulting game (which thus corresponds to a version of the Hi-Lo matching game) yields two different social welfare equilibria, both strategies can become rational for both players. However, considering our model
of social ties (with $k_{ab} = k_{ba} = 1$), as in Figure 4.9(c), shows that both players should still coordinate on the $(B, B)$ outcome, which corresponds to the unique Nash equilibrium of the resulting game. As the main result of this analysis, our model of ties clearly predicts that the outside option is irrelevant in the presence of a social tie between Alice and Bob. The players’ behavior remains the same independently of the presence of this outside option. In addition to the stability in the agents’ behavior it allows for, our model of ties appears to be more realistic than Charness & Rabin’s theory of Fairness. In fact, as shown through Section 4.5.2, convergence towards a unique social welfare equilibrium requires a high level of fairness, along with some forward induction reasoning, and a propensity to help the worst-off person over maximizing the group payoff. Assuming that human beings have bounded computational resources, our model of social ties, which clearly relies on some low level of strategic reasoning, seems definitely more adequate to solve the sort of dilemma introduced by the Entrance game.

Furthermore, one should note that both our model of social ties and the concept of team reasoning happen to make the same predictions in the context of the Entrance game, no matter whether the outside option is present or not. Such an observation therefore suggests the need for investigating other relevant game theoretic situations in future work, as a means to disentangle predictions from those theories (see, e.g., the game in Table 4.5 from Section 4.2).

### 4.10 Conclusion

In this chapter, we have proposed a game that appears to have very nice properties to investigate the behavioral effects of social ties. Indeed it creates a dilemma between maximizing self-interest and maximizing social welfare. It differs however from existing economic games from the experimental economic literature that elicit similar properties, such as the trust game, the ultimatum game, and the dictator game. Those games indeed provide situations where people’s decision may be influenced by some psychological factors such as disappointment, regret, and guilt (Geanakoplos et al.

While investigating the impact of social ties on social emotions clearly represents an interesting research orientation for future work, it is not the motivation here: the strategic structure of the Entrance game introduced in Section 6.3 does not seem to be adequate for eliciting such emotional reasoning. Moreover, a clear advantage of the Entrance game is that it is well suited to evaluate the very plausible theory of team reasoning in the context of social ties: the stronger the tie between individuals, the more they may act as members of the same group.

However, as this work is purely theoretical, it clearly requires some further experimental analysis. The next stage of this study therefore consists in testing
and evaluating the main hypotheses made in the previous section. We therefore present such a study in the next chapter.
Chapter 5

The Behavioral Effects of Social Ties

“True happiness consists not in the multitude of friends, but in the worth and choice.”
— Ben Jonson
Cynthia’s Revels (1600)

“A true friend is one soul in two bodies.”
— Aristotle

Measuring the effects of social relationships on human behavior is not new to the areas of economics and social psychology. In the past years, many experimental studies have indeed shown that people tend to cooperate more with individuals that belong to the same group than with individuals that do not (see, e.g., Brewer [1979, 1999]; Chen and Li [2009]; Tajfel and Turner [1979]; Tajfel et al. [1971]). These observations led to distinguish between the concepts of an in-group, which constitutes a social group to which an individual psychologically identifies as one of its members, and an out-group, which, by contrast, represents a group to which an individual does not identify as a member. While such in-group behavior can reasonably be induced by a wide variety of phenomena such as culture, religion, gender and race, the well known minimal group paradigm (Tajfel [1970]) suggests that even meaningless characteristics can suffice to trigger some group identification. However, despite the empirical evidence for some in-group favoritism (i.e., favoring members of one’s in-group over out-group members), one may wonder about what determines an in-group in the first place. In fact, one can easily imagine that many different levels of in-group exist, as suggested in the previous chapter: for example, an individual might, at the same time, identify as an economist, a
computer scientist, and a member of one’s favourite sport club. It is our attempt, through this chapter, to investigate the impact that varying the strength of such in-group connectedness can have on human behavior: does an individual behave similarly when interacting with a close friend, a simple acquaintance, or a perfect stranger?

For this purpose, we propose an experiment that involves two versions of the asymmetric coordination game presented in the previous chapter (i.e., with and without the outside option). In this context, we vary the strength of social ties by making players interact with partners from different in-groups (i.e., fellow members of their own sports team, members of their sports club, students of their university). The general aim of this experimental study is to verify the validity of the social ties model presented in the previous chapter.

Furthermore, we use direct questionnaires to measure more precisely social connections between each individual and his/her own sports team. In this case, we distinguish between two different types of social ties: a subjective tie measures the level with which a particular individual feels about the connectedness of his/her own team (through the individual’s own ratings about other team members), whereas an objective tie measures the level with which a team is actually close to a particular member (through the ratings of other team members about the individual). After comparing the two types of social ties (subjective and objective) with one another, we then investigate the role that each plays in determining individual behavior.

5.1 Experimental Design

Throughout this section, we present the detailed setting of our experiment that aims at measuring the effects of social ties on human behavior. More specifically, in Section 5.1.1, we specify two particular types of asymmetric coordination games that we wish to test: the Baseline game, and the Entrance game. We further introduce, in Section 5.1.2, the problem of anonymity while studying social ties, along with our proposed solution. Section 5.1.3 then presents the methodology that is followed to measure social ties within a group.

5.1.1 The coordination games

As suggested in the previous chapter (see Section 4.9), we consider two coordination games that involve two players in our experiment. The first game consists of the Baseline game, which corresponds to an asymmetric version of the well-known Battle of the Sexes game. Such a game is depicted in Figure 5.1. The asymmetric property of this game lies in the players’ payoffs: the worst outcome for Player
Player (1) is different from the worst outcome for Player (2). As shown in the previous chapter (see Section 6.3), this asymmetry characterizes the only distinction that can be made with the classical Battle of the Sexes.

\[
\begin{array}{c|cc}
& A & B \\
\hline
A & (35,5) & (0,0) \\
B & (0,0) & (15,35) \\
\end{array}
\]

Figure 5.1: Baseline game

The second game of this experimental study simply consists of an extension of the previous Baseline game with an outside option. The corresponding game, which was already introduced in the previous chapter, defines the Entrance game depicted in Figure 5.2.

\[
\begin{array}{c|cc}
& A & B \\
\hline
A & (35,5) & (0,0) \\
B & (0,0) & (15,35) \\
\end{array}
\]

Figure 5.2: Entrance game

According to the Entrance game in Figure 5.2, Player (1) has the possibility not to actually play the Baseline game by choosing Out.

Throughout this experiment, in order to incentivize subjects to play both of the above games, we consider real monetary payoffs. In this case, the currency of
the payoffs from Figures 5.1 and 5.2 is the euro: for example, if both players select A in the Baseline game, then Player (1) will get 35 € whereas Player (2) will get 5 €.

Moreover, one should note that a detailed game theoretic analysis of both of the above games was introduced in Section 4.4 from Chapter 4.

5.1.2 Preserving anonymity

The issue raised here comes from the simple property that both games that are defined in the previous section involve only two players. In fact, our aim through this work is to investigate the influence of social ties on the subjects’ behavior in the context of these simple games. However, considering social ties clearly requires knowing the identity of the individual one is tied with: how could one estimate his social relationship with someone without knowing anything about the identity of that individual? On the other hand, allowing perfect information about the identity of the subjects’ partners while playing the above games becomes problematic as it allows the possibility for some reputation effect: e.g., assuming two interacting subjects who are strongly tied with one another, each may then consider the possibility of future punishment or sharing outcomes afterwards (i.e., outside of the experiment). In order to prevent such biases, all the subjects of our experiment are therefore divided into groups such that, instead of knowing the exact identity of the partner one is matched with, one is only provided the information of which group the partner actually belongs to. For example, one may consider a group of friends, where all members are (somewhat similarly) strongly tied with one another (one may similarly consider a group of strangers). In the case of two interacting members of this group, revealing the identity of the subjects’s partner simply becomes unnecessary as both appear to be strongly tied with the group itself. This way, one can elicit social ties between interacting individuals while maintaining anonymity.

5.1.3 Measuring Social Ties

Our goal here is to measure the subjects’ social relationships with the group they belong to. In order to quantify such social ties, each subject is first required to rate his/her connection with every member of his/her group. More specifically, the question used to fill this purpose is to ask the subjects to indicate their beliefs about how they are appreciated by each other group member, based on their picture, as shown in Figure 5.3. Note that the reciprocal property implied by this question satisfies Constraint 4.1.0.1 from Section 4.1 in Chapter 4 that restricts a
social tie to be bilateral\(^1\).

![Figure 5.3: An individual’s expected tie with a group member](image)

In the context of this question, we use the four available options to define the scale of a tie according to Figure 5.4: given a group member, the strongest tie is considered whenever the subject “likes a lot” that person, whereas the weakest tie is considered whenever the subject “dislikes” that person.

![Figure 5.4: Individual tie measure](image)

Although the above question can reveal what one may call the social value of a certain subject within a certain group (e.g., the more one believes to be liked by others, the more one’s social value is important), we claim that it is not sufficient to meet our definition of a social tie with a group. Indeed, let us recall that subjects are expected to interact anonymously so that they know who they may be interacting with (i.e., a member of their group), without knowing who exactly they actually interact with. This means that a subject’s tie with an unknown group member can be reasonably interpreted as the tie with the group itself. However, as indicated in Section 4.1 from Chapter 4, a social tie is assumed to be bilateral, which therefore implies that the intensity of the relationship with a group one belongs to must be the same for every member of that group. In order to illustrate this interpretation, let us consider the following scenario: suppose that Alice is

\(^1\)The subjects were also asked to answer a similar question to that in Figure 5.3, where they were required to indicate their direct feeling about each other group member in the same fashion. As we observed that answers to both questions are strongly correlated, we chose to focus on the most restrictive case, which is depicted in Figure 5.3.
socially very close to Bob, Carol, and Daniel, while, at the same time, these three characters dislike each other (i.e., they are all extremely weakly tied with one another). Let us also suppose that Alice actually interacts with Bob. In this case, although Alice is indifferent between interacting with either character, Bob is not. Indeed, Bob is more likely to actually interact with someone he dislikes, and so Alice should take this information into account in order to make her choice. One’s tie with a group should then not only rely on one’s individual ties with other members, but it should also take into account the ties existing between every pair of members from the group. This is why we ask all subjects in our experiment to give their estimate about which member is socially tied to whom within the group they belong to. As shown through Figure 5.5, a subject is required to draw lines between members they believe are actual “friends”.

![Diagram](image.png)

Figure 5.5: An individual’s subjective estimate of others’ ties

In this case, the presence of a connection between two members is interpreted as the existence of a tie between them according to the subject who answered. Conversely, an absence of connection between two members is interpreted as a non existent tie between them according the same subject. To illustrate this with the particular example depicted in Figure 5.5, where four individuals A, B, C, and D are considered, the subject X (who answers the question) indicates his beliefs that B is only tied with D, C is only tied with A, and A is also tied with D (in
addition to being tied with $C$). Such binary measures of ties are used in order to keep the question as simple as possible to the subjects, without removing too much valuable information (as subjects are asked about others’ ties, the imperfection of such an information may indeed lead to introduce unnecessary noise through more detailed questions).

### 5.2 Experimental Procedure

In our experiments, students from Toulouse 1 University Capitole who are also members of the main university volleyball club were recruited as participants. As a preliminary phase during training sessions, every active member of this club was proposed to participate to our study. Upon acceptance, every subject was then photographed for the purpose of later measuring social ties with their own teammates (see Section 5.1.3).

The experiment itself was run in November 2011 during two training sessions. In total, 70 subjects participated, including 37 men and 33 women. As active volleyball players within the club, all subjects were divided into 9 single-sex teams: 5 teams were exclusively made of men, and 4 teams were exclusively made of women. The minimum (resp. maximum) number of subjects in a given group was 7 (resp. 9). Both training sessions can be defined as follows:

- **Session A**: 31 subjects divided into 3 male teams and 1 female team.
- **Session B**: 39 subjects divided into 2 male teams and 3 female teams.

All (male and female) teams were ranked based on their performance according to the official volleyball coach of the club. The best (i.e., most efficient, top ranked) male/female teams all belong to Session B.

It is assumed from this population of subjects that members of the same team do naturally share some social ties with one another. In fact, considering our definition of social ties from Section 4.1 in Chapter 4, these players do share some common social feature that define their social identity (e.g., they are all students at the same university, they all like sport, and particularly enjoy playing volleyball) while also having regular meaningful interactions with one another (they at least all play volleyball together for 2 hours every week).

The paper and pencil method was used all along in our experiment. At the beginning of both sessions, all subjects were asked to fill a questionnaire, which includes rather personal questions (e.g., about their hobbies, study, religious/political beliefs), as well as questions related to measuring their social tie with their own team (see Section 5.1.3 for details).

The purpose of answering this questionnaire prior to playing both games is simply to prevent the subjects’ behavior in both games to influence their ratings.
of social ties. Indeed, our goal is to measure genuine ties, which are independent of
any social context. On the other hand, it is worth mentioning that eliciting social
ties before playing the games is not a problem. In fact, it is likely that answering
the related questions may influence the subjects’ behavior in both games, which
is precisely the purpose of our experiment. Moreover, one should note that, while
measuring a social tie with an individual seems quite straightforward (either one
likes/dislikes someone or is indifferent), measuring a social tie with a group seems
rather more ambiguous. It can therefore be assumed that letting the subjects
answer these questions beforehand may lead them to become more “aware” of the
actual level with which they are close to their group.

Every subject was then asked to play both of the above Baseline and Entrance
games, according to three different types of matching processes. The detailed in-
structions of both games are described in Sections B.2 and B.3 of the Appendix.
The use of such a within-subject design is clearly justified by the reasonable as-
sumption that social ties are individual intrinsic characteristics. The purpose of
this experiment is indeed to study any possible change of behavior that may be
induced by different levels of social ties.

The three different matching processes can therefore be described as follows:

- The “university” scenario: the interaction involves a member of the volley-
ball club (i.e., a participant of this experiment) and some randomly selected
student from Toulouse 1 University Capitole who does not belong to the
volleyball club. This situation defines our control treatment, as very little
information is made available about the co-player. In this case, we assume
the existence of a very weak tie (if not absent) between the players.

- The “club” scenario: the interaction is made between two randomly selected
volleyball club members (i.e., participants of this experiment) who do not
belong to the same volleyball team. This situation illustrates the existence
of some social tie of intermediate strength between the players that mainly
relies on the limited sharing of some common social feature (e.g., enjoying
playing volleyball) and some possible few past interactions (during a usual
training session, students are indeed subjects to occasional interactions with
students that do not belong to their own team).

- The “team” scenario: the interaction is made between two randomly se-
lected members of the same volleyball team. This situation characterizes the
case with the strongest social tie existing between two subjects in this exper-

\footnote{Furthermore, this particular scenario was also independently replicated between economics
students from the Toulouse School of Economics and randomly selected students from Toulouse
1 University Capitole who are not economics students.}
iment. As said earlier, such a scenario indeed illustrates well our definition of social ties from Section 4.1 in Chapter 4.

In each of these cases, one should note that information imperfection is symmetric, that is, the type of scenario is made common knowledge among both of the players involved.

It is clear from the definitions of the above matching processes that each scenario characterizes a different level of social tie between partners, as shown through Figure 5.6.

![Figure 5.6: Quantifying social ties based on the matching process](image)

These three scenarios are then played in sequence by every subject in the context of both games, using the following meta-strategy method: for each scenario, all subjects had to indicate their decision if assigned the role of player (1), as well as their decision if assigned the role of player (2) (see Sections B.2.1 and B.3.1 in the Appendix for the detailed instructions).

Furthermore, in order to detect any possible influence the order of playing these scenarios may have on the subjects’ behavior, we distinguish two different experimental sequences in both sessions:

- In Session A: subjects first played the Entrance game before playing the Baseline game, and in each case, they considered scenarios in decreasing order of the level of social ties (i.e., starting with the “team” scenario).

- In Session B: subjects first played the Baseline game before playing the Entrance game, and in each case, they considered scenarios in increasing order of the level of social ties (i.e., starting with the “university” scenario).

It is also worth pointing out that, although each game was played repeatedly (i.e., once for every situation), each case remains a one-shot game as it is guaranteed that a subject cannot interact more than once with the same co-player in the same situation. However, note that the probability \( p \) of interacting with the same individual in both games (i.e., the Baseline and Entrance games) is \( p < 1/18000 \) in the “university” scenario, \( 1/63 < p < 1/61 \) in the “club” scenario, and \( 1/8 < p < 1/6 \) in the “team” scenario\(^1\).

Moreover, in order to elicit their beliefs about what characterizes their expected behavior in the context of both the Baseline and Entrance games played in the

\(^1\)The actual value of \( p \) in the “team” and “club” scenarios depends on the team the corresponding subject belongs to.
“university” scenario, all subjects were asked to indicate their expectations of what
decision a randomly selected student from the university would make in both roles
(i.e., both as \textit{player (1)} and as \textit{player (2)}). As shown in the instructions provided
in Section B.4 from the Appendix, subjects were also incentivized to answer care-
fully to these questions (i.e., they were offered a monetary prize whenever their
guess was accurate). The obvious purpose of these complementary questions is to
provide some extra information regarding the subjects’ way of reasoning and ratio-
nality (e.g., do people play the best response to their belief about their co-player’s
choice?).

The whole experiment lasted approximately one hour in both sessions. The
participants’ payments were distributed during the following training sessions in
December 2011. The payment method, which was clarified to all subjects be-
forehand (see Section B.1 from the Appendix for details), did then consist in
randomly drawing one role (i.e., \textit{player (1)} or \textit{player (2)}), one game (i.e., Baseline
game or Entrance game), one scenario (i.e., “university”, “club”, or “team”), and
one co-player (depending on the scenario). A subject’s payoff was therefore de-
finite according to his choice made as the selected player in the selected situation
(which corresponds to the selected scenario in the selected game), and the selected
co-player’s choice in the same situation. Each effective payment was made individ-
ually and anonymously through random draws that were performed in front of the
subject\textsuperscript{1}. All participants received the total sum of their actual earnings, which
includes a 5€ show-up fee. The mean of total payments was 19.03€ (standard
deviation of 12.21€, with a maximum of 40€ and a minimum of 5€).

5.3 Results

This section presents descriptive statistics reporting the various elicited behavior
throughout our experiment. First, in Section 5.3.1, we describe the players’ ob-
served behavior in both the Baseline game and the Entrance game, for various
levels of social ties. One should note that, in this case, the strength of a social
tie is artificially controlled by changing the type of a subject’s game partner. We
will therefore consider three distinct levels of such social ties corresponding to the
three scenarios defined in Section 5.2 (i.e., “team”, “club”, or “university”). Then
in Section 5.3.2, we propose to refine this analysis by considering subjective and
objective measures of each participant’s social tie with his/her respective team: in
a particular group (e.g., a volleyball team), the members may indeed be more or
less tied with each other (e.g., some members may be friends while others are not).

\textsuperscript{1}The random selection of the co-player was made through some code name in order to preserve
anonymity between subjects.
We therefore investigate the effects of these two different types of ties on behavior in the context of the above games.

5.3.1 General observations

Prior to analysing the observed behavior in both the Baseline game and the Entrance game, it is worth indicating that we observe no effect regarding the order of playing either game: behavior of subjects from Session A can indeed not be distinguished from that of subjects from Session B (see Section 5.2 for details about the experimental design).

5.3.1.1 Behavior in the Baseline game

Table 5.1 represents the players’ resulting behavior in the Baseline game, depending on whether the corresponding co-player is a teammate, a club member, or a university student. Table 5.1 also includes the \( p \) values related to the Wilcoxon signed rank tests for similarity of the subjects’ behavior in various scenarios. Note that, in all following tables in this chapter, only \( p \) values lower than 0.2 are displayed. \( p \) values larger than 0.2 are classified as not significant (n.s.). The first observation one can make from Table 5.1 is that the subjects are torn between choosing A and B when assigned the role of Player (1) in the presence of some weak tie with Player (2) (i.e., in the “university” scenario). This randomizing behavior may simply be the direct consequence of the conflict existing between Player (1)’s own preferences, and the group’s welfare: Player (1) indeed prefers the \((A, A)\) outcome while \((B, B)\) is clearly better for the group (and for Player (2)). Moreover, note that the elicited behavior, which largely differs from the optimal mixed strategy (i.e., play \( A \) with probability \( 7/8 \)), suggests that the subjects are well aware of this conflict, and can therefore hardly choose between satisfying their self-interest and satisfying the welfare of the group.

<table>
<thead>
<tr>
<th>Players</th>
<th>Matching types</th>
<th>Wilcoxon signed rank test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>1 (70 obs.)</td>
<td>67%</td>
<td>57%</td>
</tr>
<tr>
<td>2 (70 obs.)</td>
<td>75%</td>
<td>76%</td>
</tr>
</tbody>
</table>

Table 5.1: Choosing B in the Baseline game for each player in each type of matching

One can also observe that subjects tend to favor option B significantly more often whenever the social tie with Player (2) increases. We can indeed reject
the null hypothesis that Player (1)’s behavior when interacting with a university student is the same as Player (1)’s behavior when interacting with a teammate ($p < 0.003$). This means that, as Player (1), increasing one’s social tie with Player (2) allows to accept giving up some of one’s own payoff in order to favor the group made of both players. In other words, the existence of a (strong) social tie between the players simply allows to reveal the existence of a focal point to Player (1), which corresponds to the unique best outcome for the group. Furthermore, note that the elicited behavior then goes further apart from the optimal mixed strategy as the tie increases, which indicates the presence of some sufficiently strong incentive to satisfy the welfare of the group$^1$.

Similarly, when assigned the role of Player (2), the subjects clearly favor playing $B$ in all types of interactions, which is a major difference with Player (1)’s behavior from Table 5.1. In this case, Player (2)’s observed behavior is close to the optimal mixed strategy (i.e., playing $B$ with probability $7/10$), which may support the subjects’ intention to satisfy their self-interest. This result is however not very surprising because, unlike Player (1), Player (2)’s preferences perfectly match that of the group (i.e., there is no conflict between Player (2)’s individual preferences and the group’s welfare). As a consequence of facing no dilemma, the subjects need not care about the welfare of the group when assigned the role of Player (2). However, one should note that Player (2)’s choice does not vary significantly with an increase of the social tie’s strength. This clearly indicates that Player (2) does not even take into consideration his/her corresponding tie with Player (1) in order to make a choice. This result is rather surprising because, by anticipating Player (1) to choose $B$ more often in the presence of a stronger tie, a purely self-regarding rational Player (2) would also choose $B$ more often. Therefore, Player (2)’s unchanging behavior suggests that the subjects may actually not be so purely self-regarding after all.

5.3.1.2 Behavior in the Entrance game

Tables 5.2 and 5.3 represent the players’ resulting behavior in both stages of the Entrance game, depending on whether the co-player is a teammate, a club member, or a university student. More specifically, Table 5.2 depicts Player (1)’s choice between $In$ and $Out$ during the first stage of the game. Table 5.3 similarly depicts both players’ behavior in the second stage of the game (i.e., the Entrance subgame), in the hypothetical case that the second stage were reached (through Player (1) playing $In$ in the first stage)$^2$. Tables 5.2 and 5.3 also include the $p$ values related to

$^1$Also note from Table 5.1 that an intermediate level of social ties (i.e., through the “club” scenario) induces some existing but less significant change in behavior.

$^2$Table 5.3 therefore includes Player (1)’s counterfactual choice in the second stage: if choosing $Out$ in the first stage, what would Player (1) have played in the subgame had he chosen $In$
the Wilcoxon signed rank tests for similarity of the subjects’ behavior in various scenarios. However, as such statistical tests cannot be performed over Player (1)’s whole strategy space in the Entrance game (i.e., Player (1) has four discrete choices: (In, A), (In, B), (Out, A), and (Out, B)), we simply provide the observed behavior in details through Figure 5.7, which can be read as follows: according to Figure 5.7(c), among the 42% of subjects who chose In in the first stage of the game, 52% then played A in the subgame. However, in the same context, 43% of the subjects who chose Out first would have played A in the subgame had they chosen In first.

<table>
<thead>
<tr>
<th>Matching types</th>
<th>Wilcoxon signed rank test (p values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>62%</td>
<td>53%</td>
</tr>
<tr>
<td>team vs. university</td>
<td>team vs. club</td>
</tr>
<tr>
<td>0.004</td>
<td>0.083</td>
</tr>
</tbody>
</table>

Table 5.2: Player (1) choosing In in the first stage of the Entrance game (70 obs.)

Concerning Player (1)’s elicited behavior in the presence of some weak tie with Player (2) (i.e., in the “university” scenario), the first observation one can make from Table 5.2 is that the subjects play Out more often (58%). Moreover, Table 5.3 shows that in this context, the subjects are torn between choosing either strategy A or B in the subgame. More precisely, Figure 5.7(c) indicates that this observation is particularly true among the subjects who played In in the first stage. Such a result is rather surprising as it does not suggest any strong common belief in each other’s rationality. Indeed, as shown in the analysis of the Entrance game from Chapter 4 (see Section 4.4), if Player (1) believes in Player (2)’s rationality and that Player (2) believes in Player (1)’s rationality, then Player (1)’s only rational move is to play (In, A), which corresponds to the forward induction reasoning. Moreover, it appears that considering the weaker assumption of bounded rationality does also not suffice to explain all this elicited behavior: no matter what Player (1) believes about Player (2)’s future move, playing (In, B) can never be selected as a rational self-regarding move. Yet, Figure 5.7(c) shows that 20% of the subjects actually selected strategy (In, B) in this context. As a means to provide a realistic interpretation of this observation, one can observe that the outside option in the first stage of the Entrance game is not relevant to the subjects’ decision as Player (1) in the subgame. One can indeed not reject the null hypothesis that, in the context of the “university” scenario, Player (1)’s choice in the Entrance subgame (i.e., after choosing first In) is the same as Player (1)’s choice in the Baseline game. This result therefore suggests that, right after instead?
playing In, Player (1) tends to consider the subgame as a new independent game (i.e., Player (1) then forgets about the previous outside option).

<table>
<thead>
<tr>
<th>Players</th>
<th>Matching types</th>
<th>Wilcoxon signed rank test (p values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>1 (70 obs.)</td>
<td>63%</td>
<td>55%</td>
</tr>
<tr>
<td>2 (70 obs.)</td>
<td>77%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Table 5.3: Choosing B in the second stage of the Entrance subgame

Considering Player (2)’s choice in the context of weak ties (i.e., in the “university” scenario), Table 5.3 similarly suggests that the subjects tend to play the optimal mixed strategy in the subgame (i.e., playing B with probability 7/10). Following this interpretation, Player (2) may act rationally in response of believing that Player (1) also acts rationally in the subgame only (i.e., without considering the outside option). It is also worth noting that Player (2)’s observed behavior from Table 5.3 does largely differ from the mixed strategy equilibrium in the entire Entrance game (i.e., playing B with probability 3/7 as shown in Section 4.4 from Chapter 4). However, it appears that, unlike Player (1), Player (2)’s behavior is somewhat affected by Player (1)’s outside option. One can indeed reject the null hypothesis that, in the context of the “university” scenario, Player (2)’s choice is the same in the Entrance subgame and in the Baseline game (p < 0.008). As a result, the fact that Player (2) chooses A significantly more often in the Entrance game than in the Baseline game suggests that Player (2) is sensible to some forward induction reasoning (see Section 4.4 from Chapter 4).

Consequently, both players’ elicited behavior in the “university” scenario clearly illustrates the failure of the principle of individual rationality. Furthermore, this analysis suggests the existence of some group-oriented behavior: the design of the Entrance game indeed allows for the dominant outcome (In, B; B) to be the best outcome for the group made of both Player (1) and Player (2).

Moreover, focusing on other scenarios that consider the presence of stronger social ties between the players allows to reinforce this hypothesis: one can indeed observe from Figure 5.7 that subjects (as Player (1)) tend to favor option (In, B) more often whenever the social tie level between players increases. According to Table 5.2, we can indeed reject the null hypothesis that Player (1)’s initial choice between In and Out is the same when interacting with a university student as when interacting with a teammate (p < 0.005). Similarly, Table 5.3 allows to reject the null hypothesis that Player (1)’s behavior in the subgame (independently of initially playing In or Out) is the same in the “team” scenario and in the
“university” scenario ($p < 0.04$). Thus, this result indicates that stronger social ties induce Player (1) to play $B$ significantly more often than $A$ in the subgame. Figures 5.7(a) and 5.7(c) further indicate that Player (1)’s changing behavior in the subgame (through choosing $B$ more often than $A$) generally follows a change in the initial move (through choosing $In$ more often than $Out$). In other words, Player (1) generally plays strategy ($In, B$) significantly more often when a stronger social tie is involved with Player (2).

Focusing on the effect of social ties on Player (2)’s behavior in the Entrance game, Table 5.3 also reveals some significant difference: Player (2) is more likely to play $B$ in the presence of a stronger tie with Player (1). One can reject the null hypothesis that Player (2)’s behavior is the same when interacting with a university student as when interacting with a teammate ($p < 0.05$). Note however that, although Player (1) cannot be rational to play strategy ($In, B$), both actions for Player (2) (i.e., $A$ and $B$) are rationalizable: Player (2) should rationally select $B$ (resp. $A$) as a best response of believing that Player (1) will also play $B$ (resp. $A$). In other words, this means that Player 2’s observed behavior can always be
justified by some rationality assumption.

Let us now continue the previous analysis regarding the influence of the outside option on both players’ behavior. In the presence of strong social ties (i.e., in the “team” scenario), Player (1)’s behavior in the Entrance subgame (after choosing first In) is not significantly different from Player (1)’s behavior in the Baseline game alone. Similarly, in the same context, Player (2)’s behavior is also not significantly different in the Entrance game as compared to the Baseline game. These results therefore indicate that, although Player (1)’s behavior in the Entrance subgame is independent of the outside option (i.e., no matter the strength of the tie with Player (2)), the strategic effect of the outside option on Player (2)’s behavior (in the “university scenario”, as shown earlier) is removed through the introduction of some stronger social tie with Player (1). In other words, in the presence of some sufficiently important social tie, Player (2) tends to ignore the outside option and simply reach the most profitable outcome for the group.

5.3.2 Refining social ties

5.3.2.1 Behavioral effects of subjective social ties

As shown through the previous sections, every subject in our experiment was asked to provide subjective information about how tied they were with each of their teammate. Our aim in this section, is to aggregate these answers in a way that best characterizes all subjects’ closeness to their team. We therefore choose to define social ties as subjective values, such that a subject’s relationship with a team depends exclusively on that subject’s own beliefs. Let us also recall from Section 5.1.3 that one’s tie with a group not only depends on one’s closeness to every other member, but it also depends on every other member’s closeness to each other. In other words, assuming all uncertainty was removed, every member of the same team should then be similarly tied with that team. Therefore, one advantage of this method is that it allows to later compare every teammate’s subjective tie with each other, as an indication of how well the team members actually know each other. In order to measure such social ties, let us first focus on the level of individual ties each subject can have about some teammate.

Based on our experimental data, and as a matter of simplicity, we choose to strictly consider boolean ties one suspects to exist between two individuals (i.e., whether two teammates are socially tied with each other or not). For this purpose, let us first translate one’s belief about one’s own social ties with every other team member (see Figure 5.3 in Section 5.1.3) into boolean values. To do so, we consider the following interpretation, which appears to be the most restrictive:

- if an individual A believes that a teammate B “likes a lot” A, then A is considered to be socially tied with B;
• if an individual $A$ believes that a teammate $B$ either “likes”, “dislikes”, or is “indifferent” towards $A$, then $A$ is considered to be not socially tied with $B$;

Given a group of individuals $G$, let us then denote $N_i$ the number of such individual ties that some individual $i$ (such that $i \in G$) has within group $G$ (according to $i$).

In addition to these individual ties with others, we also consider each individual’s subjective estimate of others’ ties, as shown in Figure 5.4 from Section 5.1.3. Let us then similarly define $N_{-i}$ to represent the estimated number of individual ties within group $G$ (according to $i$’s beliefs) that do not involve $i$. For example, assuming individual $B$ provides the information depicted in Figure 5.4, $B$ then believes that there are two individual ties within the team that do not involve $B$, so in this case: $N_{-B} = 2$.

We then simply describe one’s level of social tie with a group $G$ as one’s expected belief that any two members from $G$ can be tied with one another. Let us define an individual $i$’s subjective social tie $k_i^s$ with a group $G$ (assuming that $i \in G$) as the level with which $i$ believes to be closely connected to $G$. Formally, this measure is characterized as follows:

$$k_i^s = \frac{N_i + N_{-i}}{N}$$

where $N$ corresponds to the maximum number of individual ties in $G$:

$$N = \frac{|G| \times (|G| - 1)}{2}$$

Applying this measure to our population of volleyball players, the distribution of subjective social tie values is depicted in Figure 5.8. One can observe from this graph that most participants do not feel strongly tied with their own team (e.g., 67% of the subjects have social tie value lower than or equal to 0.4).

We find an average social tie value of 0.35 with a standard deviation of 0.23. However, one can also note that the entire range of social tie values is covered: the minimum and maximum tie values are respectively 0.035 and 1 in this population.

Following the previous analysis from Section 5.3.1.1 that relied on a measure of social ties controlled artificially through varying the type of game partner, we here refine this study by considering the effects of the individuals’ subjective social ties on some fixed group. For this purpose, we propose to split the entire set of subjects from our experiment into two equally sized sets (High and Low) based on their subjective social tie values $k_i^s$: each dataset represents the subpopulation consisting of the 35 subjects with the highest/lowest social tie values $k_i^s$. The resulting behavior in the Baseline game based on this distribution is shown in Table 5.4. In this case, note that Table 5.4 also includes the $p$ values related to the
Mann-Whitney tests for similarity of behavior between these two populations of subjects (i.e., $High_s$ vs. $Low_s$). One can then observe from the “team” matching that the subjective tie value with one’s team seems to affect Player (1)’s behavior. Such a result is confirmed by performing a logistic regression, which shows that the higher the players’ subjective social tie, the more likely they are to play $B$ when assigned the role of Player (1) in the “team” scenario ($p < 0.005$, no. of obs. = 70). However, running a similar regression regarding behavior as Player (2) (also in the “team” scenario) shows no evidence of any such correlation with social ties.

<table>
<thead>
<tr>
<th>Subj. ties $k^*_i$</th>
<th>Players</th>
<th>Matching types</th>
<th>Wilcoxon signed rank test (p values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>$High_s$</td>
<td>1</td>
<td>82%</td>
<td>71%</td>
</tr>
<tr>
<td>(35 obs.)</td>
<td>2</td>
<td>76%</td>
<td>83%</td>
</tr>
<tr>
<td>$Low_s$</td>
<td>1</td>
<td>51%</td>
<td>43%</td>
</tr>
<tr>
<td>(35 obs.)</td>
<td>2</td>
<td>74%</td>
<td>69%</td>
</tr>
<tr>
<td>$High_s$ vs. $Low_s$ (p values - Player 1)</td>
<td>0.017</td>
<td>0.016</td>
<td>0.057</td>
</tr>
<tr>
<td>$High_s$ vs. $Low_s$ (p values - Player 2)</td>
<td>n.s.</td>
<td>0.166</td>
<td>0.182</td>
</tr>
</tbody>
</table>

Table 5.4: Choosing $B$ in the Baseline game based on subjective social ties $k^*_i$
Another interesting observation one can make from Table 5.4 is that focusing on the individuals with lower social ties (i.e., Low\_s) still reveals some significant difference across scenarios. In fact, we can reject the null hypothesis that, under the assumption of low subjective social tie values, Player (1)’s behavior is the same in the “team” scenario and in the “university” scenario (p < 0.1). This result therefore suggests that the group itself has some independent effect on its members’ behavior, even when those members are very weakly tied with each other. Note, however, that this observation does again not duplicate to Player (2) whose behavior does not change significantly across scenarios.

Let us now similarly refine the previous analysis from Section 5.3.1.2 to investigate the effects of the individuals’ subjective social ties with their team in the context of the Entrance game. As in Section 5.3.1.2 (see Tables 5.2 and 5.3), Tables 5.5 and 5.6 depict the players’ resulting behavior in both stages of the Entrance game based on this distribution. Note that Tables 5.5 and 5.6 also include results of the Mann-Whitney tests as in Figure 5.4. Furthermore, we provide the observed behavior of Player (1) in more details through Figures 5.9 and 5.10, which can be read as in Figure 5.7 from Section 5.3.1.2.

<table>
<thead>
<tr>
<th>Subjective ties ( k_s^i )</th>
<th>Matching types</th>
<th>Wilcoxon signed rank test (p values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>High_s</td>
<td>63%</td>
<td>49%</td>
</tr>
<tr>
<td>Low_s</td>
<td>62%</td>
<td>57%</td>
</tr>
<tr>
<td>High_s vs. Low_s ( (p \text{ values}) )</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Table 5.5: Player (1) choosing \textit{In} in the Entrance game based on subjective social ties \( k_s^i \) (35 obs.)

One can first see from Table 5.5 that subjective social ties have apparently no influence on selecting the outside option as Player (1) in the context of the “team” scenario. In fact, no significant correlation can be found regarding the subjects’ level of subjective ties with the team (i.e., high or low) and their choice between \textit{In}/\textit{Out} in this case. However, concerning Player (1)’s behavior in the subgame (independently of choosing \textit{In}/\textit{Out} first), Table 5.6 suggests that individuals with higher subjective ties with the team tend to favor choosing \textit{B} rather than choosing \textit{A}. This observation is confirmed by performing a logistic regression between one’s subjective tie value with a team, and one’s behavior in the “team” scenario of the Entrance subgame (p < 0.1, no. of obs. = 70).

However, it is worth pointing out that, as for the Baseline game, some signifi-
Table 5.6: Choosing $B$ in the Entrance subgame based on subjective social ties $k^s_i$.

cant differences in behavior still exist across scenarios for both types of (high/low) subjective social ties. Indeed, as shown in Tables 5.5 and 5.6, one can observe some significant change in Player (1)’s behavior in the presence of high subjective ties with their team. More surprisingly, these results show similar effects in the context of low subjective tie values. In fact, the subjects tend to distinguish weak social ties within their team (i.e., the “team” scenario), and weak social ties within a large group of unknown individuals (i.e., the “university” scenario): the subjects from $Low_s$ actually favor playing $In$ significantly more often in the “team” scenario than in the “university” scenario ($p < 0.09$ according to Table 5.5). This observation duplicates to the second stage of the game where they play $B$ (independently of initially choosing $In$ or $out$) significantly more often in the “team” scenario ($p < 0.1$ according to Table 5.6).

Furthermore, concerning the subjects’ behavior in the role of Player (2), one can also observe a correlation between one’s level of subjective social tie with a team and one’s behavior in any type of scenario: the higher Player (2)’s subjective tie with his/her team, the more likely Player (2) is to select $B$ (even if Player (2) interacts with a stranger, as in the “university” scenario). Note, however, that such a correlation is not very significative, as shown in Table 5.6 ($p > 0.18$).

5.3.2.2 Behavioral effects of objective social ties

In contrast with the analysis from the previous section, we here verify the relevance of an objective measure of social ties in determining one’s social behavior.

While subjective ties measure how an individual feels connected to his/her own team (see previous section), objective ties aim at measuring how the team feels...
Figure 5.9: Elicited behavior for Player (1) with high subjective social ties $k_i^s$ (35 obs.)

connected (i.e., through its members) to a particular individual. In this case, one may value an individual $i$’s objective social tie with a group by only considering the average subjective social tie (as measured in the previous section) of every other member of that group (excluding $i$). In this case, one can define an individual $i$’s objective social tie $k_i^o$ with a group $G$ (assuming that $i \in G$) as the level with which $i$ is closely connected to that group according to other members of $G$. Formally, this measure is characterized as follows:

$$k_i^o = \frac{\sum_{j \in G, j \neq i} k_j^s}{(|G| - 1)}$$

Again, applying this measure of objective ties to our population of volleyball players leads to the corresponding distribution displayed in Figure 5.11. In this case, we find an average social tie value of 0.35 with a standard deviation of 0.15. The minimum and maximum tie values are respectively 0.13 and 0.64 in this population.
Let us now analyse the influence of such objective social ties on the subjects’ behavior in both of our economic games. For this purpose, we again propose to split the entire set of subjects from our experiment into two equally sized sets (High_o and Low_o) based on their objective social tie values $k_o$: each dataset represents the subpopulation consisting of the 35 subjects with the highest/lowest social tie values $k_o$. The resulting behavior based on this distribution is shown in Table 5.7.

One can clearly observe from Table 5.7 that, in the context of the Baseline game, there exists some correlation that reveals the following rather surprising observation: the lower one’s objective social tie value, the more one is cooperative and play $B$ as Player (2) in the Baseline game. Note that this apparent correlation is confirmed by performing a logistic regression explaining one’s behavior in the “team” scenario in terms of one’s objective tie value ($p < 0.05$, no. of obs. = 70). This observation duplicates to Player (1), even though it does not represent a significant correlation.

However, one may wonder whether this result also applies to the behavior

---

Figure 5.10: Elicited behavior for Player (1) with low subjective social ties $k_i^s$ (35 obs.)
Figure 5.11: Distribution of objective social tie values

<table>
<thead>
<tr>
<th>Obj. ties $k_i^o$</th>
<th>Players</th>
<th>Matching types</th>
<th>Wilcoxon signed rank test $(p$ values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>$High_o$</td>
<td>1</td>
<td>61%</td>
<td>54%</td>
</tr>
<tr>
<td>(35 obs.)</td>
<td>2</td>
<td>68%</td>
<td>83%</td>
</tr>
<tr>
<td>$Low_o$</td>
<td>1</td>
<td>71%</td>
<td>60%</td>
</tr>
<tr>
<td>(35 obs.)</td>
<td>2</td>
<td>83%</td>
<td>69%</td>
</tr>
<tr>
<td>$High_o$ vs. $Low_o$</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>(p values - Player 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$High_o$ vs. $Low_o$</td>
<td>0.103</td>
<td>0.166</td>
<td>0.182</td>
</tr>
<tr>
<td>(p values - Player 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7: Choosing $B$ in the Baseline game based on objective social ties $k_i^o$

The results can be observed from Table 5.8 that such objective social ties have a significant effect on Player (1) selecting In/Out in the first stage of the game. This result is confirmed by performing a logistic regression, which suggests that subjects with low objective ties with their team are more likely to play In as Player (1) ($p < 0.07$, no. of obs. = 70). However, while this observation confirms
Table 5.8: Player (1) choosing In in the Entrance game based on objective social ties $k_i^o$ (35 obs.)

<table>
<thead>
<tr>
<th>Objective ties $k_i^o$</th>
<th>Matching types</th>
<th>Wilcoxon signed rank test $(p$ values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>$High_o$</td>
<td>53%</td>
<td>41%</td>
</tr>
<tr>
<td>$Low_o$</td>
<td>71%</td>
<td>65%</td>
</tr>
<tr>
<td>$High_o$ vs. $Low_o$</td>
<td>0.087</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Table 5.9: Choosing $B$ in the Entrance subgame based on objective social ties $k_i^o$

<table>
<thead>
<tr>
<th>Obj. ties $k_i^o$</th>
<th>Players</th>
<th>Matching types</th>
<th>Wilcoxon signed rank test $(p$ values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>$High_o$</td>
<td>1</td>
<td>59%</td>
<td>55%</td>
</tr>
<tr>
<td>(35 obs.)</td>
<td>2</td>
<td>79%</td>
<td>67%</td>
</tr>
<tr>
<td>$Low_o$</td>
<td>1</td>
<td>71%</td>
<td>55%</td>
</tr>
<tr>
<td>(35 obs.)</td>
<td>2</td>
<td>74%</td>
<td>65%</td>
</tr>
<tr>
<td>$High_o$ vs. $Low_o$ (p values - Player 1)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>$High_o$ vs. $Low_o$ (p values - Player 2)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

As a result, while this empirical analysis suggests that objective social ties have some limited impact on behavior as compared to subjective ties (see previous section), it also indicates that both types of tie measures are relevant to determine one’s propensity to cooperate in the context of both the Baseline game and the Entrance game. We therefore attempt to investigate an intuitive combination of both concepts through the next section.
5.3.2.3 Behavioral effects of underestimating and overestimating social ties

We have shown, in the previous sections, that both subjective and objective measures of social ties matter to specify behavior in the Baseline game as well as the Entrance game. In this section, we propose to analyse the behavioral effects of a combination of both concepts. For this purpose, we simply consider the difference value between one’s subjective tie and one’s objective tie, i.e., $k_s^i - k_o^i$. In fact, the previous analyses from Sections 5.3.2.1 and 5.3.2.2 suggested that high subjective ties and low objective ties independently tend to promote some fair and cooperative behavior. We therefore investigate whether overestimating one’s social tie with one’s team (i.e., $k_s^i > k_o^i$) does have some significant effect on one’s behavior.

In our population of volleyball players, the corresponding distribution of the difference between tie measures (i.e., $k_s^i - k_o^i$) is displayed in Figure 5.14. In this case, we find an average difference value of 0 with a standard deviation of 0.2. The minimum and maximum difference values are respectively -0.49 and 0.49 in this
population.

One should note from Figure 5.14 that the population of subjects is basically divided into two categories: a group of individuals who underestimate their social ties (i.e., 37 individuals \( k_s^i < k_o^i \)) and a group of individuals who overestimate their social ties (i.e., 33 individuals \( k_s^i > k_o^i \)). We therefore propose to look at the elicited behavior in each of these groups.

Focusing on the Baseline game, the resulting behavior based on this categorization is shown in Table 5.10.

One should note from Table 5.10 that there is a very significant change in behavior between subjects from the two different groups (i.e., \( k_s^i > k_o^i \) and \( k_s^i < k_o^i \)) when assigned either the role of Player (1) or Player (2).

Moreover, looking at the Entrance game similarly reveals some strong changes in behavior, as shown through Tables 5.11 and 5.12.

Although there does not exist a very significant change in behavior between both groups (i.e., \( k_s^i > k_o^i \) and \( k_s^i < k_o^i \)) in the first stage of the Entrance game (see...
Table 5.10: Choosing $B$ in the Baseline game based on $k_i^s$ and $k_i^o$

<table>
<thead>
<tr>
<th>Social tie measures</th>
<th>Players</th>
<th>Matching types</th>
<th>Wilcoxon signed rank test $(p$ values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>team</td>
<td>club</td>
</tr>
<tr>
<td>$k_i^s &gt; k_i^o$</td>
<td>1</td>
<td>91%</td>
<td>82%</td>
</tr>
<tr>
<td>(33 obs.)</td>
<td>2</td>
<td>88%</td>
<td>88%</td>
</tr>
<tr>
<td>$k_i^s &lt; k_i^o$</td>
<td>1</td>
<td>44%</td>
<td>35%</td>
</tr>
<tr>
<td>(37 obs.)</td>
<td>2</td>
<td>64%</td>
<td>65%</td>
</tr>
<tr>
<td>$k_i^s &gt; k_i^o$ vs. $k_i^s &lt; k_i^o$ $(p$ values - Player 1)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>$k_i^s &gt; k_i^o$ vs. $k_i^s &lt; k_i^o$ $(p$ values - Player 2)</td>
<td>0.014</td>
<td>0.026</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Table 5.11), it is clear from Table 5.12 that behavior in the subgame follows the same interpretation as in the previous Baseline game: when overestimating their social ties, subjects are much more likely to play $B$ in either role (i.e., Player (1) or Player (2)).

Moreover, another main observation that can be made through this analysis of both of the above games concerns the subjects’ behavior across scenarios. In fact, one can observe from Tables 5.10, 5.11 and 5.12 that, unlike subjects who overestimate social ties, subjects who underestimate ties do not appear to modify
their behavior much across the three different scenarios. This observation can be particularly emphasized by looking at Player (1)'s behavior in the Entrance game. More precisely, while Figures 5.15 clearly shows that Player (1)'s behavior when overestimating the social tie with Player (2) significantly changes across scenarios, Figure 5.16 indicates that, when underestimating the social tie with Player (2), Player (1) only changes behavior in the subgame if selecting the outside option first (i.e., Out), which appears to be irrelevant as the subgame will not be reached in this case (i.e., it represents strictly hypothetical behavior).

Furthermore, it is also worth noting from Tables 5.10 and 5.12 that the effects of one’s overestimated social tie with a team go beyond the context of the “team”
scenario. In fact, the more one overvalues the tie with a given team, the more one plays B in the Baseline game and the Entrance subgame when assigned the role of Player (1) in the “university” scenario, even though Player (1) shares no apparent tie with Player (2) in this case. This observation therefore suggests that one’s subjective tie value with a given team relies to some extent on some genuine notion of “fairness”: individuals who overvalue their social tie with a team are more prone to express a “fair” behavior when interacting with some stranger. On the other hand, one should note that this observation does not duplicate to Player (2)’s behavior.

5.4 Discussion

While our results from Section 5.3.1 show that social ties have a clear effect on behavior in the context of the Baseline game and the Entrance game from Section 5.1.1, the refined analysis from Section 5.3.2 confirms this observation and adds the following extra results:
• Both subjective and objective measures of social ties matter to eventually affect behavior: results from Sections 5.3.2.1 and 5.3.2.2 indicate that high subjective ties and low objective ties allow to improve coordination in the above games.

• Overestimating social ties promotes some cooperative behavior. As shown through Section 5.3.2.3, subjects who overestimate social ties coordinate more than those who do not. On the other hand, underestimating social ties leads the subjects to barely differentiate between playing with a teammate and playing with a stranger (e.g., a university student).

It is our attempt through this section to extend the previous study with some relevant analyses in order to better interpret the behavior that we observed. In Section 5.4.1, we investigate the subjects’ efficiency in both of the above games and compare these observations with theoretical predictions. Section 5.4.2 provides experimental data that we collected about the subjects’ beliefs during the experiment. Finally, in Section 5.4.3, we consider our experiment from a different
angle, which provides evidence for some fair behavior.

5.4.1 Efficiency

Let us now analyse the efficiency of the subjects’ behavior in both games.

Table 5.13 depicts each player’s expected payoff in the Baseline game according to various types of scenarios (either “team”, “club”, or “university”) and three groups of subjects based on their subjective and objective social ties with their own team: we consider the group of individuals who overestimate their social ties (i.e., $k^s_i > k^o_i$), the group of individuals who underestimate their social tie (i.e., $k^s_i < k^o_i$), as well as the group of all individuals. We further assume that the subjects strictly interact with co-players of the same group (e.g., subjects overestimating ties interact with a co-player who also overestimate ties).

One can observe from Table 5.13 that, although $Player (1)$’s behavior is significantly changing across scenarios (as shown in Table 5.10), the corresponding expected payoff remains stable (for any type of subjective tie).

<table>
<thead>
<tr>
<th>Social tie measures</th>
<th>Players</th>
<th>Matching types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>team</td>
</tr>
<tr>
<td>$k^s_i &gt; k^o_i$</td>
<td>1</td>
<td>12€</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28€</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>40€</td>
</tr>
<tr>
<td></td>
<td>Difference</td>
<td>16€</td>
</tr>
<tr>
<td>$k^s_i &lt; k^o_i$</td>
<td>1</td>
<td>11€</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>11€</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>22€</td>
</tr>
<tr>
<td></td>
<td>Difference</td>
<td>0€</td>
</tr>
<tr>
<td>All</td>
<td>1</td>
<td>11€</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>18€</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>29€</td>
</tr>
<tr>
<td></td>
<td>Difference</td>
<td>7€</td>
</tr>
</tbody>
</table>

Table 5.13: Expected payoffs in the Baseline game based on $k^s_i$ and $k^o_i$ (rounded to full euros)

In fact, the change in $Player (1)$’s behavior appears to mainly benefit $Player (2)$, whose expected payoff is clearly increasing with increasing ties (see the “team” scenario). Note however that the expected payoff is very similar for both players in the case of underestimated social ties when $k^s_i < k^o_i$ group (i.e., in the “club” and “university” scenarios in Table 5.13). In other words, underestimating ties
seems to simply promote equality in the context of our Baseline game. Moreover, note that, in this case, subjects manage to reach a nearly optimal egalitarian outcome since the best expected payoff that both players can equally obtain through mixed strategies is $11.6 \epsilon$ (i.e., when both players choose $B$ with probability $3 - \sqrt{6} \approx 0.55$). This means that increasing one player’s expected payoff above this value has for immediate effect to decrease the other’s expected payoff and therefore reduce equality. This also implies that overestimating social ties could not possibly improve the social welfare of the group without reducing equality. As a consequence, if social ties were to promote equality, then they should induce individuals to play more randomly when assigned to either role (i.e., choose $B$ with probability $\approx 0.55$), which is clearly not what we observe in our experiment. In fact, Table 5.13 shows that stronger social ties when $k_i^s > k_i^o$ instead improve the pair’s expected payoff\(^1\), which could be maximized if both players were to choose $B$ with probability 1 (the pair’s payoff would then be of $50 \epsilon$).

The unfortunate consequence of improving the group’s expected payoff through stronger overestimated social ties, as shown in Table 5.13, is thus the unavoidable increase of inequality between co-players. This result therefore indicates that Player (1)’s behavior in the presence of (strong) social ties aims at only promoting utilitarianism (by improving the group’s total welfare) at the cost of sacrificing equality. Such an observation can be justified by the attractiveness of reaching the $(B, B)$ outcome, which not only maximizes the pair’s total payoff, but also maximizes the payoff of the worst-off individual. Furthermore, note that playing $(B, B)$ appears to be more beneficial for both players than any randomization of strategies as suggested by the most equitable expected payoff (no individual can indeed get more than $15 \epsilon$ through randomizing strategies).

Let us now similarly consider the various payoffs each individual can expect in the Entrance game, based on the observed behavior from Section 5.3.2.3. According to Table 5.14, one can first note that increasing the level of social ties promotes Player (2)’s expected payoff without changing Player (1)’s whenever $k_i^s > k_i^o$. Such a result appears as a direct consequence of Player (1) selecting $(In, B)$ more often, which turns out to allow Player (2)’s expected payoff to increase at the expense of decreasing Player (1)’s. However, one should note that such a behavior is not purely altruistic as it is more beneficial to Player (2) than it is costly to Player (1). In other words, social ties simply allow to improve the social welfare of the pair while also reducing the expected inequality between both players. Note from Table 5.14 that, while overestimating social ties favors Player (2) (in addition to promoting group efficiency), underestimating ties tends to favor Player (1).

Unlike for the Baseline game, it appears that maximizing the efficiency of the pair somehow coincides with maximizing the equitable expected payoff in the Entrance game. However, one should note that such a behavior is not purely altruistic as it is more beneficial to Player (2) than it is costly to Player (1).
Table 5.14: Expected payoffs in the Entrance game based on $k_s^i$ and $k_o^i$

<table>
<thead>
<tr>
<th>Social tie measures</th>
<th>Players</th>
<th>Matching types</th>
<th>team</th>
<th>club</th>
<th>university</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_s^i &gt; k_o^i$</td>
<td>1</td>
<td>15 €</td>
<td>15 €</td>
<td>17 €</td>
<td></td>
</tr>
<tr>
<td>(33 obs.)</td>
<td>2</td>
<td>22 €</td>
<td>14 €</td>
<td>12 €</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>37 €</td>
<td>29 €</td>
<td>29 €</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>7 €</td>
<td>1 €</td>
<td>5 €</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_s^i &lt; k_o^i$</td>
<td>1</td>
<td>15 €</td>
<td>15 €</td>
<td>16 €</td>
<td></td>
</tr>
<tr>
<td>(37 obs.)</td>
<td>2</td>
<td>9 €</td>
<td>11 €</td>
<td>10 €</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>24 €</td>
<td>26 €</td>
<td>26 €</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>6 €</td>
<td>4 €</td>
<td>6 €</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>1</td>
<td>15 €</td>
<td>15 €</td>
<td>16 €</td>
<td></td>
</tr>
<tr>
<td>(70 obs.)</td>
<td>2</td>
<td>15 €</td>
<td>12 €</td>
<td>11 €</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>30 €</td>
<td>27 €</td>
<td>27 €</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>0 €</td>
<td>3 €</td>
<td>5 €</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

trance game: the unique way to maximize the welfare of the group is for Player (1) to always choose strategy (In, B) while Player (2) always selects B (i.e., with probability 1 for both players). In this case, Player (1) obtains 15 € and Player (2) obtains 35 € (and the pair therefore gets 50 €). On the other hand, the maximum equitable expected payoff of 18.33 € for both players (and therefore 36.66 € for the pair) can be uniquely reached through Player (1) randomizing between Out (with probability 2/3) and (In, B) (with probability 1/3) while Player (2) always selects B (i.e., with probability 1). However, as our previous analysis in Section 5.3.2.3 indicates that increasing social ties leads to a decreasing probability of Player (1) selecting Out, it suggests that subjects do not follow this mixed strategy. As a consequence of this analysis, one can conclude that subjects of our experiment act more as utilitarians (i.e., aiming at maximizing the pair’s total expected payoff) than as egalitarians (i.e., aiming at maximizing the pair’s equitable expected payoff) in both the Baseline game and the Entrance game.

5.4.2 Relationship between beliefs and behavior

In this section, we provide a comparative analysis between the subjects’ behavior and their beliefs in the context of both of the previous games. As shown in Section B.4 from the Appendix, all participants to our experiments were asked to estimate their quantitative beliefs about how students from the university would generally behave (i.e., in the “university” scenario). Table 5.15 depicts, for each player’s strategy in both the Baseline game and the Entrance game, the rate of individuals
actually making that choice along with the average quantitative belief that this choice will be selected in general. These average beliefs are measured through the answers of all participants to the questions from Section B.4 in the Appendix. More detailed results can be found in Section B.5 from the Appendix, where the graphs depict each player’s beliefs about the generally expected behavior in each game.

<table>
<thead>
<tr>
<th>Games</th>
<th>Players</th>
<th>Strategies</th>
<th>“university” scenario</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Average behavior</td>
<td>Average beliefs</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>1</td>
<td>B</td>
<td>49%</td>
<td>41%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>B</td>
<td>73%</td>
<td>72%</td>
<td></td>
</tr>
<tr>
<td>Entrance</td>
<td>1</td>
<td>(In, A)</td>
<td>22%</td>
<td>34%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(In, B)</td>
<td>20%</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Out</td>
<td>58%</td>
<td>42%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>B if In</td>
<td>53%</td>
<td>59%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B if Out</td>
<td>47%</td>
<td>41%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>B</td>
<td>64%</td>
<td>64%</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.15: Choices versus beliefs in both games

One can observe from Table 5.15 that the average beliefs are surprisingly close to the actual average behavior, which suggests that subjects behave as they believe others would behave in the same situation. Such an interpretation however questions the existence of rational self-regarding behavior in the context of our experiment. In order to clarify this argument, let us consider the well known concept of epistemic rationality ([Aumann 1995]), which states that a player is epistemically rational if and only if he does not “knowingly select a strategy that yields him less than he could have gotten with a different strategy”. Following this concept, we express, in Table 5.16, every rational choice that is compatible with every possible type of quantitative belief that can be held by subjects in both roles (i.e., Player (1) and Player (2)). In this case, we abbreviate \( Pr_i(X) \) to denote individual \( i \)’s quantitative belief that proposition \( X \) holds (\( Pr_i(X) \in [0,1] \)). One should note that the conditions on beliefs in Table 5.16 are determined through the theoretical equilibrium solutions in mixed strategies presented in Section 4.4 from Chapter 4. As an example, one can indeed state that, in the Baseline game, if Player (1) believes that Player (2) will choose to uniformly randomize strategies (i.e., \( Pr_1(“2 selects A”) = Pr_1(“2 selects B”) = 0.5 \)), then Player (1)’s only rational move is to select A. However, in the same Baseline game, if Player (1) believes that Player (2) will play according to the equilibrium solution in mixed strategies (i.e., \( Pr_1(“2 selects B”) = 0.7 \)), Player (1) is rational to perform either
of the available options (i.e., A or B).

<table>
<thead>
<tr>
<th>Games</th>
<th>Players</th>
<th>Conditions on beliefs</th>
<th>Rational choices</th>
<th>Observed rationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1</td>
<td>$Pr_1(\text{“}2 \text{ selects } B\text{”}) = 7/10$</td>
<td>$A/B$</td>
<td>58%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Pr_1(\text{“}2 \text{ selects } B\text{”}) &gt; 7/10$</td>
<td>$B$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Pr_1(\text{“}2 \text{ selects } B\text{”}) &lt; 7/10$</td>
<td>$A$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$Pr_2(\text{“}1 \text{ selects } B\text{”}) = 1/8$</td>
<td>$A/B$</td>
<td>54%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Pr_2(\text{“}1 \text{ selects } B\text{”}) &gt; 1/8$</td>
<td>$B$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Pr_2(\text{“}1 \text{ selects } B\text{”}) &lt; 1/8$</td>
<td>$A$</td>
<td></td>
</tr>
<tr>
<td>Entrance</td>
<td>1</td>
<td>$Pr_1(\text{“}2 \text{ selects } B\text{”}) = 3/7$</td>
<td>$\text{Out}/(\text{In}, A)$</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Pr_1(\text{“}2 \text{ selects } B\text{”}) &gt; 3/7$</td>
<td>$\text{Out}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Pr_1(\text{“}2 \text{ selects } B\text{”}) &lt; 3/7$</td>
<td>$(\text{In}, A)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$Pr_2(\text{“}1 \text{ selects } B \text{ if } \text{In}\text{”}) = 1/8$</td>
<td>$A/B$</td>
<td>59%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Pr_2(\text{“}1 \text{ selects } B \text{ if } \text{In}\text{”}) &gt; 1/8$</td>
<td>$B$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Pr_2(\text{“}1 \text{ selects } B \text{ if } \text{In}\text{”}) &lt; 1/8$</td>
<td>$A$</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.16: Rationality in both games (“university” scenario)

Furthermore, for every role (i.e., Player (1) and Player (2)) in each game (i.e., Baseline and Entrance games), Table 5.16 also depicts the rate of rational behavior that is actually observed throughout our experiment, based on the subjects’ selected actions and beliefs. As a result, Table 5.16 does clearly not support a strongly rational behavior. For example, in the context of the Entrance game, we observe that only half of the subjects did actually behave rationally when assigned the role of Player (1). Moreover, one should note that the data from our experiment does not allow to detect any correlation between one’s choice of action as a given player and one’s quantitative belief about the co-player’s behavior in either game. Such results therefore confirm our analyses from the previous sections (see Section 5.3), which did already not indicate any strong rationality effect.

However, following our previous interpretation of Table 5.15, data from our experiment provides empirical evidence for an existing connection between the subjects’ beliefs and their actions. In fact, performing a logistic regression reveals a clear correlation between one’s action and one’s quantitative belief about what others would do in the same situation (for every player in each game: $p < 0.005$, no. of obs. = 70): for example, in the Baseline game, the higher one’s quantitative belief about Player (1) choosing $B \text{ in general}$, the more likely one is to similarly select $B$ when assigned the role of Player (1). Such an observation therefore suggests that, when asked to play either of the above coordination games, subjects favor some sort of normative thinking over the generally assumed principle of rationality. Such a result can easily be justified by the lack of any salient
optimal solution in both games: as shown in the previous chapter (see Section 4.4),
while there is no unique Nash equilibrium in both coordination games, the forward
induction solution in the Entrance game may not even be followed by perfectly
rational players who do not believe that their partner is also perfectly rational. As
a result of this inability to perform any rational action, the subjects are therefore
led to follow some simple rule that consists in doing what they believe others would
do in their position. Note, however, that this result does not mean that people
are not self-regarding. In fact, although this study suggests the absence of any
strategic thinking, it also provides evidence for some individualistic behavior: in
the absence of any strong tie, people generally aim at reaching the best outcome
for themselves, hoping that their co-player will act accordingly.

5.4.3 Behind the veil of ignorance

In this section, we interpret our experiment through a different viewpoint, which
refers to the concept of the “original position”, as introduced by Rawls in his theory
of justice (Rawls [1971]). According to Rawls, the original position defines an ideal
situation in which agents have to make a collective agreement about what is fair
for every individual. The main distinguishing feature of this original position is
the so-called “veil of ignorance”, behind which every individual is assumed to be
deprived of all knowledge about his personal identity and characteristics so that
he imagines himself to possibly be in any player’s position.

This concept of the original position is clearly relevant to our experiment. In
fact, Sections B.2.1 and B.3.1 in the Appendix show that the meta-strategy method
used in our design allows the subjects to make their choice behind such a veil of
ignorance: in the context of both the Baseline game and the Entrance game, all
participants were asked to make their decision as both players, not knowing which
role would eventually be effective. However, in order to analyse the corresponding
behavior under this assumption, one needs to interpret both games differently.
Indeed, in the original position, a given strategy corresponds to a combination of
Player (1) and Player (2)’s actions. As a result, the original games are simply
transformed into symmetric two player games where the two players, say X and
Y, have the same strategy space: for example, in the Baseline game, strategy
(A, B) is interpreted as playing A if assigned the role of Player (1), and playing B
if assigned the role of Player (2). In this case, a player’s outcome is determined as
follows: if X chooses strategy (B, A) and Player Y chooses strategy (A, B), then
assigning X to the role of Player (1) (and therefore Y to the role of Player (2))
leads X to obtain 15 € (i.e., the effective outcome becomes (B, B)) while assigning
X to the role of Player (2) (and therefore Y to the role of Player (1)) leads X to
obtain 5 € (i.e., the effective outcome becomes (A, A)).

In order to determine X’s payoff in such a transformed game, Harsanyi argues
in Harsanyi [1986] that one should assign equal probabilities to the fact of playing as either player. Following this assumption, X’s expected payoff is then simply obtained as the simple average of the two possible assignments (e.g., in the above example, X gets \((15 + 5)/2 = 10\) \(\varepsilon\)). Note that, in the context of the Baseline game, Players X and Y both have four different strategies. The corresponding payoff matrix for the whole transformed Baseline game can be found in Table 5.17, where only Player X’s expected payoff is depicted (as the game is symmetric, Player Y’s payoff can be obtained by simply inverting X with Y in Table 5.17).

\[
\begin{array}{cccc}
  & (A, A) & (A, B) & (B, A) & (B, B) \\
(A, A) & 20 & 2.5 & 17.5 & 0 \\
(A, B) & 17.5 & 0 & 35 & 17.5 \\
(B, A) & 2.5 & 10 & 0 & 7.5 \\
(B, B) & 0 & 7.5 & 17.5 & 25 \\
\end{array}
\]

Table 5.17: Payoffs à la Harsanyi for row Player X in the transformed Baseline game

Note that, according to Table 5.17, the only strategies that are evolutionary stable are \((A, A)\) and \((B, B)\).

However, one should note that Player X’s payoff in the transformed Baseline game could be determined differently. In fact, Rawls alternatively argues, in Rawls [1971], that one should seek to maximize the utility of the worst-off player when playing behind the veil of ignorance. For example, if X chooses strategy \((B, A)\) and Player Y chooses strategy \((A, B)\), then X’s payoff becomes \(\min(15, 5) = 5\) \(\varepsilon\) (instead of \((15 + 5)/2 = 10\) \(\varepsilon\) according to Harsanyi’s view presented above). The corresponding payoff matrix for the new transformed Baseline game can be found in Table 5.18, where again only Player X’s payoff is depicted.

\[
\begin{array}{cccc}
  & (A, A) & (A, B) & (B, A) & (B, B) \\
(A, A) & 5 & 0 & 0 & 0 \\
(A, B) & 0 & 0 & 35 & 0 \\
(B, A) & 0 & 5 & 0 & 0 \\
(B, B) & 0 & 0 & 0 & 15 \\
\end{array}
\]

Table 5.18: Payoffs à la Rawls for row Player X in the transformed Baseline game

According to both Tables 5.17 and 5.18, one can note that the Baseline game has three different pure Nash equilibria when played behind the veil of ignorance:

- both players selecting \((A, A)\);
• both players selecting \((B, B)\);
• one player choosing \((A, B)\) while the other chooses \((B, A)\).

In this case, Table 5.19 depicts the actually elicited behavior in our experiment, according to the various types of matching.

<table>
<thead>
<tr>
<th>Strategies</th>
<th>Matching types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>team</td>
</tr>
<tr>
<td>((A, A))</td>
<td>14%</td>
</tr>
<tr>
<td>((A, B))</td>
<td>19%</td>
</tr>
<tr>
<td>((B, A))</td>
<td>10%</td>
</tr>
<tr>
<td>((B, B))</td>
<td>57%</td>
</tr>
</tbody>
</table>

Table 5.19: Behavior in the original position of the Baseline game

The first main observation from Table 5.19 lies in the actual dominant strategy being \((A, B)\) in the “university” scenario, as if the subjects were then trying to take advantage of each other (by expecting others to select \((B, A)\) as suggested by the above equilibrium). However, note that, according to the previous section, subjects who play \((A, B)\) in this game tend to believe that others would also play \((A, B)\), which is not consistent with the principle of rationality (in both Figures 5.17 and 5.18, this yields a payoff of 0). This observation therefore confirms the presence of some self-regarding interest without any strategic thinking. Furthermore, Table 5.19 also shows that social ties allow to reduce this individualistic behavior by significantly increasing the rate of selecting \((B, B)\) (57% in the “team” scenario, against only 34% in the “university” scenario).

In order to interpret this result, let us consider Binmore’s theory from Binmore [1994, 1998, 2005], which claims that, when placed behind the veil of ignorance, two individuals naturally tend to share the same preferences. In this case, Binmore argues that, when making a choice in the original position, since the players have a common aim, any strategic thinking becomes irrelevant. As a result, the players will simply “agree” to choose whichever solution maximizes the commonly shared preferences.

Applied to the above Baseline game, this interpretation predicts that, if the players make their decision behind the veil of ignorance, then they will select strategy \((B, B)\). In fact, as shown in Tables 5.17 and 5.18, both players selecting \((B, B)\) is better than both players selecting any other strategy. Following this theory, our observations from Table 5.19 therefore suggest that social ties promote

---

1In Binmore [1994, 1998, 2005], Binmore defines one’s preferences in the original position as one’s empathetic preferences (the next chapter provides a more detailed discussion about such empathetic preferences).
the use of the original position as a coordination device in the context of the Baseline game.

Let us now perform a similar analysis of the Entrance game being played behind the veil of ignorance. Following Harsanyi’s assumption that the players assign equal probabilities to the fact of playing as either player (i.e., Player (1) or Player (2)), the corresponding payoff matrix for the transformed Entrance game can be found in Table 5.20, where only Player X’s payoff is depicted. One should note that, in this case, both have six different available strategies.

\[
\begin{array}{cccccc}
\text{Player X} & (\text{In, A; A}) & (\text{In, A; B}) & (\text{In, B; A}) & (\text{In, B; B}) & (\text{Out; A}) & (\text{Out; B}) \\
\hline
(\text{In, A; A}) & 20 & 2.5 & 17.5 & 0 & 22.5 & 5 \\
(\text{In, A; B}) & 17.5 & 0 & 35 & 17.5 & 22.5 & 5 \\
(\text{In, B; A}) & 2.5 & 10 & 0 & 7.5 & 5 & 12.5 \\
(\text{In, B; B}) & 0 & 7.5 & 17.5 & 25 & 5 & 12.5 \\
(\text{Out; A}) & 12.5 & 12.5 & 10 & 10 & 15 & 15 \\
(\text{Out; B}) & 10 & 10 & 27.5 & 27.5 & 15 & 15 \\
\end{array}
\]

Table 5.20: Payoffs à la Harsanyi for row Player X in the transformed Entrance game

According to Table 5.20, the transformed Entrance game has three different pure Nash equilibria:

- both players selecting (In, A; A);
- both players selecting (Out; B);
- one player choosing (In, A; B) while the other chooses (Out; A).

Note that, out of these solutions, only strategy (In, A; A) is evolutionary stable.

However, as for the Baseline game, the player’s payoff in the transformed Entrance game could be determined according to Rawls’ view. In this case, the corresponding payoff matrix for the new transformed Entrance game can be found in Table 5.21, where again only Player X’s payoff is depicted.

According to Table 5.21, the new transformed Entrance game has the following different pure Nash equilibria (note that it differs from the previous interpretation in Table 5.20):

- both players selecting (In, A; A);

\footnote{For simplicity reasons, we voluntarily omit counterfactual strategies (i.e., (Out, A; ·) and (Out, B; ·)) that are irrelevant to this analysis.}
Table 5.21: Payoffs à la Rawls for row Player X in the transformed Entrance game

- both players selecting (Out; ·);
- one player choosing (In, A; ·) while the other chooses (Out; A).
- one player choosing (In, B; ·) while the other chooses (Out; B).

In this case, the subjects’ actually elicited behavior in our experiment is depicted in Table 5.22.

<table>
<thead>
<tr>
<th>Strategies</th>
<th>Matching types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>team</td>
</tr>
<tr>
<td>(In, A; A)</td>
<td>12%</td>
</tr>
<tr>
<td>(In, A; B)</td>
<td>11%</td>
</tr>
<tr>
<td>(In, B; A)</td>
<td>4%</td>
</tr>
<tr>
<td>(In, B; B)</td>
<td>35%</td>
</tr>
<tr>
<td>(Out; A)</td>
<td>7%</td>
</tr>
<tr>
<td>(Out; B)</td>
<td>31%</td>
</tr>
</tbody>
</table>

Table 5.22: Behavior in the original position of the Entrance game

The first observation one can make regarding interactions with weak ties (i.e., in the “university” scenario) is that only few people do actually select the most optimal and stable strategy (In, A; A). However, Table 5.22 also shows that social ties significantly increase the rate of selecting (In, B; B) (35% in the “team” scenario, against only 13% in the “university” scenario).

As for the previous Baseline game, applying Binmore’s theory to the above Entrance game predicts that, if the players make their decision behind the veil of ignorance, then they will select strategy (In, B; B). In fact, as shown in Tables 5.20 and 5.21, both players selecting strategy (In, B; B) is better than both players selecting any other strategy.
As a main result, this analysis of both the Baseline game and the Entrance game when played behind the veil of ignorance provides evidence showing that social ties induce people to use the original position to make their decision, which has for effect to remove the rate of miscoordination in both games.

Moreover, it is worth mentioning that this theory is consistent with the experimental data about the subjects’ beliefs from Section 5.4.2. In fact, this analysis shows that social ties do not only reinforce the use of normative thinking, they also allow to replace individualistic behavior with fair behavior: indeed, behind the veil of ignorance, the only problem an individual faces is to figure out what is the right thing to do for the group.

5.5 Conclusion

In this chapter, we have presented an economic experiment of the coordination games previously introduced in Chapter 4. As a main result of this study, we provided some empirical evidence revealing that stronger social ties help people coordinate and promote the welfare of the group. More specifically, our observations support the proposed model of social ties from Chapter 4 (see Section 4.2), while rejecting other relevant theories of social preferences (see Section 4.5). Furthermore, we have shown that social ties allow to promote a sense of fairness in the context of the above games, thereby driving people to make their decision as if they were unable to distinguish their own identity from that of their partner (cf. Rawls’ original position). This analysis therefore suggests that social ties can be used as a device that leads the players to bridge the gap between individualistic behavior (when decreasing social ties) and fair behavior (when increasing social ties), as specified by our model of social ties from the previous chapter.

Moreover, this experimental analysis shows that one’s behavior is affected by both one’s expected (subjective) social tie with another individual, as well as one’s actual (objective) social tie. More precisely, we observed that optimistic people who overestimate the strength of their social ties tend to behave more fairly by cooperating with others, whereas pessimistic people who underestimate the value of their social ties do not identify with any group and therefore fail to coordinate in the above games. Such a result clearly suggests that being self-regarding does not always pay off, especially in the context of coordination games such as those introduced here. In fact, in this type of social interactions, it is in one’s own best interest to be socially tied with other individuals, as a means to eventually coordinate and maximize profits.

It is also worth mentioning that this study also indicates that there exists no absolute “zero” on the continuous scale of a social tie between two individuals: we indeed observe that when interacting with university students (which can be
assumed to be almost perfect strangers), some subjects already choose to cooperate as if they were socially tied with those people. In this case, our definition of social ties from the previous chapter (see Section 4.1) provides a sound explanation for this behavior: the fact that they share the property of (1) being a student, and (2) from the same university may indeed create a sufficiently strong social connection that induces cooperation.

However, while this simple study validates our model of social ties in the context of simple two player games, it does clearly not allow to explain how social ties affect individuals’ behavior in larger societies that involve more complex types of social interactions (i.e., with possibly more than two individuals). We therefore attempt to extend this study by generalizing our theoretical model of social ties in the next chapter.
Chapter 6
Towards Collaborative Societies

“In the great non-zero-sum games of history, if you’re part of the problem, then you’ll likely be a victim of the solution.”
— Robert Wright
Nonzero (2000)

“It has been more profitable for us to bind together in the wrong direction than to be alone in the right one.”
— Nassim Nicholas Taleb
The Black Swan (2007)

Many everyday tasks require individuals to act collectively and to coordinate for the pursuit of a common goal. Examples include musicians from an orchestra who need to act together in a specific way in order to play some intended symphony, or players from a soccer team who have to coordinate with each other so that they can eventually score a goal. Even among other tasks that could be done by a single individual, many would be achieved more effectively through teamwork, e.g., painting a house, carrying an heavy object. A common property of all these situations is that each individual in the team acts as a team member and intends to do his part in the joint action of the team. Collective intentionality has been, through the last decades, a central topic in social philosophy (see, e.g., Bratman [1992, 1993]; Searle [1990]; Tuomela and Miller [1988]) as well as in economics (see, e.g., Bacharach [1999, 2006]; Sugden [1993, 2000]).

The general aim of this chapter is to use game theory to bridge the gap between individually egoistic behavior and social cooperation in the context of strategic interactions. For this purpose, we provide an analysis of a central theory from the economics literature, which explains how agents, either human or artificial, can manage to solve coordination problems in the context of a joint activity:
Bacharach’s theory of team reasoning. After discussing the limitations of this theory in modeling some intuitive social behavior, we present a generalization of the model of social ties introduced in Chapter 4, and show the various advantages such a model offers compared to Bacharach’s theory, especially in the context of social interactions in which different competing groups may coexist.

More precisely, the study we present here focuses on the central concept of group identification, which often allows to uncover the ‘focal point’ required to solve a given coordination problem (Schelling [1960]). For example, consider two individuals, say Alice and her partner Bob who have to paint their room together. Let us assume that they can paint it either in blue or in green and that they both prefer to paint it in green. Moreover, each of them is responsible for buying a tin of paint. Which color of paint will Alice and Bob buy? The solution in which each of them buys a tin of green paint becomes the ‘focal point’ of their coordination problem, because it satisfies their common goal, i.e., to paint the room in green. Therefore, if Alice and Bob reason as team members, they will likely coordinate on the joint action of buying two tins of green paint, which is the most useful for the whole team.

However, we argue, through this chapter, that such binary group identification (i.e., either one identifies with a group or not), as it is assumed by Bacharach’ theory of team reasoning, is not sufficient to model real-life situations where some individuals identify with a given group up to a certain degree: as an example, one may indeed be torn between identifying with a group of friends and identifying with a group of family members. We therefore show how our proposed theory of social ties allows to efficiently deal with this kind of dilemmas.

### 6.1 Problems of co-operation

This section is devoted to present the most well-known type of two player games that involves co-operation, as an illustration of the failure of classical economic theories to predict the behavior actually followed by human beings. A general form of this sort of games is depicted in Figure 6.1, in which case each player (Alice and Bob as respectively the row player and the column player) can either cooperate (i.e., play $C$) or defect (i.e., play $D$). The main obvious characteristics of this game is that both players can then obtain the same payoffs if and only if they manage to coordinate with each other (i.e., they each get $x$ or $y$ depending on whether they both play $C$ or $D$).

However, another important property of the game in Figure 6.1 is that the value of the payoff $z$ can provide some incentives for each individual to defect. In fact, Table 6.1 shows that the game in Figure 6.1 can refer to three different well known games from the literature, depending on the payoff $z$. 

147
In the particular case of the Hi-Lo matching game (i.e., when \( z = 0 \)), one can detect no incentive for each individual to defect. However, as already shown in Chapter 2, the interesting observation is that traditional game theory interprets it as a social dilemma where no prediction can be made under the assumption of rationality (the game indeed yields two distinct Nash equilibria, i.e., \((C, C)\) and \((D, D)\)), although it is shown in Bacharach [2006]; Colman et al. [2008] that people largely coordinate on the most rewarding outcome for both players, i.e., \((C, C)\).

Concerning the Stag-Hunt game (i.e., when \( y \leq z < x \)), also known as the “assurance game” or the “trust dilemma”, although it appears to have the same theoretical properties as the Hi-Lo game (the game then yields the same two Nash equilibria), it introduces some incentives for defecting: in fact, playing \( C \) involves a risk of losing if the other player defects, whereas playing \( D \) ensures a payoff of at least \( y > 0 \). However, as for the Hi-Lo game, experimental evidence has shown that people largely cooperate by playing \( C \) even if it is risky to do so (Van Huyck et al. [1990]).

Finally, the situation characterized by the Prisoner’s Dilemma (i.e., when \( x \leq z \)) increases further the above incentives so that cooperating now becomes strictly dominated by defecting for both players (i.e., \((D, D)\) then becomes the unique Nash equilibrium). Yet, various experimental studies have also shown a non-negligible cooperation rate (of reaching the \((C, C)\) outcome) in such scenarios, which varies between 30-40% (see, e.g., Shafir and Tversky [1992]).

In order to formally analyze such social dilemmas in more details, let us first consider a classical strategic game structure \( G = \langle \text{Agt}, \{S_i | i \in \text{Agt}\}, \{U_i | i \in \text{Agt}\} \rangle \)
as defined in Definition 2.1 from Chapter 2 (see Section 2.2.1). We denote $S_J = \prod_{i \in J} S_i$ the set of joint strategies performed by every group of agents $J \in 2^{Agt}$ where $2^{Agt} = 2^{Agt} \setminus \{\emptyset\}$. For notational convenience, throughout this chapter we write $S$ instead of $S_{Agt}$.

Moreover, for every agent $i \in Agt$, we define $Group(i)$ to be the set of all groups from $Agt$ that include $i$:

$$Group(i) = \{J \in 2^{Agt} | i \in J\}$$

One might then wonder whether current theories of social preferences proposed in the field of behavioral economics (see Fehr and Schmidt [2006] for an overview of these theories) are able to explain empirical evidences of mutual cooperation in the previous types of games (i.e., coordinating on $(C,C)$). The main idea of these theories consist in ‘transforming’ the agents’ utility in the original game on the basis of some social feature such as altruism, inequity aversion or fairness in order to obtain a new game in which equilibria can be computed using classical solution concepts (e.g., Nash equilibrium).

The theory of social preferences that seems to be most relevant to the above games relies on the notion of fairness. In Charness and Rabin [2002], Charness & Rabin indeed propose a specific form of social preferences they call quasi-maximin preferences. In their model, a collective payoff is computed by means of a social welfare function which corresponds to a weighted combination of Rawls’ maximin and of the utilitarian welfare function (i.e., summation of individual payoffs) (see [Charness and Rabin, 2002, p. 851]).

Formally, for every strategy profile $s \in S$, according to Charness & Rabin’s fairness model, the utility function of player $i \in \{Alice, Bob\}$ is given by:

$$U^F_i(s) = (1 - \lambda) \cdot U_i(s) + \lambda \cdot SW_i(s)$$

where $\lambda \in [0, 1]$ and $SW_i(s)$ defines the social welfare function as follows:

$$SW_i(s) = \delta \cdot \min_{j \in Agt} U_j(s) + (1 - \delta) \cdot \sum_{j \in Agt} U_j(s)$$

where $\delta \in [0, 1]$.

The two parameters $\lambda$ and $\delta$ can be interpreted as follows: $\lambda$ measures how much player $i$ cares about pursuing the social welfare versus his own self-interest. In the social welfare function, $\delta$ measures the degree of concern for helping the worst-off person versus maximizing the total social surplus. Setting $\delta = 1$ corresponds to the pure “maximin” principle (or “Rawlsian” criterion), while setting $\delta = 0$ corresponds to pure utilitarianism (aiming at global efficiency).

However, although such a model allows to explain why people choose to cooperate in the context of the Prisoner’s Dilemma, it remains indecisive in the
case of the Hi-Lo game or the Stag-Hunt game. In fact, in those games, such a model appears to be of no help because, for any possible value of \( \delta \) and \( \lambda \), the two strategy profiles \((C, C)\) and \((D, D)\) remain Nash equilibria in the resulting transformed game. As a consequence, similarly to the original game considering pure self-regarding agents, no prediction can be made.

Since current theories of social preferences are not sufficient to explain the behavior observed in situations like the Hi-Lo game and the Stag-Hunt game, we provide through the next sections some alternative theories that allow to correctly model and explain such observations\(^1\). As those theories refer to the concept of collective utility, let us first extend our previous formalization of a strategic game from Definition 2.1 in Chapter 2 (see Section 2.2.1) so that it incorporates the group utility function.

**Definition 6.1 (Game with Group Utility)** A strategic game with group utility is a tuple \(G = \langle \text{Agt}, \{S_i|i \in \text{Agt}\}, \{U_J|J \in 2^{\text{Agt}}\} \rangle\) where:

- \(\text{Agt} = \{1, \ldots, n\}\) is the set of agents;
- \(S_i\) defines the set of strategies for agent \(i\);
- \(U_J : \prod_{i \in \text{Agt}} S_i \to \mathbb{R}\) is a total payoff function mapping every strategy profile to some real number for some team \(J\).

For notational convenience, throughout this chapter, we write \(U_i\) instead of \(U_{\{i\}}\).

While nothing specifies how to compute each group utility in Definition 6.1, one can define those in terms of individual payoffs. As some examples, the following principles can be considered, which are well known to be the most realistic.

For every \(J \in 2^{\text{Agt}}\) and every \(s \in S\), let us define:

- classical utilitarianism (global efficiency): \(U_J(s) = \sum_{i \in J} U_i(s)\)
- Rawlsian criterion of fairness (maximin principle): \(U_J(s) = \min_{i \in J} U_i(s)\)

Note, however, that such a collective utility function may however be defined differently.

\(^1\)Other theories of social preferences such as altruism (Levine [1998]), reciprocity (Rabin [1993b]), and inequity aversion (Fehr and Schmidt [2006]) are not discussed here because they simply reduce the game from Figure 6.1 to another type of Hi-Lo matching game.
6.2 Team reasoning

In this section, we consider the well-known concept of team reasoning, which was first introduced by Bacharach in Bacharach [1999] in order to explain observed behavior in the game from Section 6.1. In order to interpret the intuitive behavior in such games, some theorists have indeed proposed to incorporate new modes of reasoning into game theory. For instance, starting from the work of Gilbert (Gilbert [1989]) and Reagan (Regan [1980]), some economists and logicians (e.g., Lorini [2011]) have studied team reasoning (also called we-mode reasoning) as an alternative to the best-response reasoning assumed in traditional game theory, also called I-mode reasoning (Bacharach [1999]; Colman et al. [2008]; Sugden [2000]).

Team-directed reasoning is the kind of reasoning that people use when they perceive themselves as acting as members of a group or team (Sugden [2000]). That is, when an agent $i$ engages in team reasoning, he identifies himself as a member of a group of agents $J$ and conceives $J$ as a unit of agency acting as a single entity in pursuit of some collective objective. A team reasoning player acts for the interest of his group by identifying a strategy profile that maximizes the collective payoff of the group, and then, if the maximizing strategy profile is unique, by choosing the action that forms a component of this strategy profile.

6.2.1 Definitions

In order to formally illustrate Bacharach’s original theory introduced in Bacharach [1999], let us extend the strategic game defined by Definition 6.1 in Section 6.1, and consider what Bacharach calls an unreliable team interaction (UTI) structure, that is, a game structure in which there is a probability that a given player identifies with a team and chooses the action which maximizes the team benefit (i.e., the player plays in the we-mode), and another probability that the player is a self-interested agent who tries to maximize his own benefit (i.e., the player plays in the I-mode). In this sense, the interaction is “unreliable” because there is no certainty that a player will reason and act as a team member.

Definition 6.2 (Unreliable Team Interaction) An unreliable team interaction structure is a tuple $UTI = \langle Agt, \{S_i | i \in Agt\}, \{U_J | J \in 2^{Agt}\}, \{\Omega_i | i \in Agt\}\rangle$ where:

- $\langle Agt, \{S_i | i \in Agt\}, \{U_J | J \in 2^{Agt}\}\rangle$ is a strategic game with group utility according to Definition 6.1;

1 As a matter of simplicity in our model, we assume the absence of any outside signal observable by the players, since it does not appear to be relevant to our study (such signals are considered in Bacharach’s original UTI structure from Bacharach [1999]).
• $\Omega_i$ is a probability distribution over the set of all possible groups $\text{Group}(i)$.

Intuitively, for any agent $i \in \text{Agt}$, $\text{Group}(i)$ is the set of groups agent $i$ may identify with. For every $J \in \text{Group}(i)$, $\Omega_i(J)$ is the probability that agent $i$ identifies with team $J$ (i.e., the probability that agent $i$ reasons and acts as a member of team $J$). In the definition of an UTI structure it is implicitly assumed that an agent $i$ identifies with a unique team at a given moment which can be either the singleton $\{i\}$, which corresponds to agent $i$ playing in the I-mode, or some set of agents $J \in \text{Group}(i)$ such that $|J| > 1$, which corresponds to agent $i$ playing in the we-mode.

The set of group identification states $\text{Groups}$ is defined as:

$$\text{Groups} = \prod_{i \in \text{Agt}} \text{Group}(i)$$

Elements of $\text{Groups}$ are denoted by $g, g', \ldots$. Given some group identification state $g \in \text{Groups}$, we write $g_i$ (such that $g_i \in \text{Group}(i)$) to denote the element of $g$ corresponding to agent $i$ (i.e., the group agent $i$ identifies with according to $g$). The probability distribution $\Omega$ over the set $\text{Groups}$ is defined as $\Omega = \prod_{i \in \text{Agt}} \Omega_i$.

Given some $(J_1, \ldots, J_n) \in \text{Groups}$, $\Omega((J_1, \ldots, J_n))$ is the probability that “agent 1 identifies with team $J_1$ and agent 2 identifies with team $J_2$ and... and agent $n$ identifies with team $J_n$”.

In Bacharach [1999], Bacharach further introduces the notion of a protocol, which consists in specifying, for every agent $i$ and every group $J$ agent $i$ may identify with, the action that $i$ should play when identifying with $J$. Formally, a protocol $\alpha$ is a function mapping every agent $i \in \text{Agt}$ and every team $J \in \text{Group}(i)$ agent $i$ may identify with to a strategy $s_i \in S_i$ for agent $i$. Given $J \in 2^{\text{Agt}}$ and $i \in J$, $\alpha(i, J) \in S_i$ therefore defines agent $i$’s strategy when identifying with the group $J$. The set of all such protocols is denoted by $\Delta$. Moreover, given a protocol $\alpha \in \Delta$ and a group $J \in 2^{\text{Agt}}$, we write $\alpha^J = \prod_{i \in J} \alpha(i, J)$ to denote the strategy of group $J$ specified by protocol $\alpha$ (i.e., $\alpha^J \in S_J$). In this case, $\alpha(i)$ is nothing but the action that agent $i$ should play according to protocol $\alpha$, when reasoning in the I-mode.

Given the set of agents $\text{Agt} = \{1, \ldots, n\}$ and the probability distribution $\Omega$ on the set of group identification states $\text{Groups}$, one can then express the expected value of a given protocol $\alpha \in \Delta$ for some team $J \in 2^{\text{Agt}}$:

$$\text{EV}_J(\alpha) = \sum_{g \in \text{Groups}} \Omega(g) \cdot U_J(\alpha(1,g_1), \ldots, \alpha(n,g_n))$$

Intuitively, $\text{EV}_J(\alpha)$ measures how much utility team $J$ can expect from the protocol $\alpha$. 

152
Furthermore, given two protocols $\alpha$ and $\beta$, and a group $J \in 2^{Agt^*}$, we write $\alpha^J \cdot \beta^{-J}$ to denote the protocol such that: (1) for all $i \in J$, $\alpha^J \cdot \beta^{-J}(i, J) = \alpha(i, J)$, and (2) for all $H \in 2^{Agt^*}$ such that $H \neq J$, and for all $i \in H$, $\alpha^J \cdot \beta^{-J}(i, H) = \beta(i, H)$.

This allows us to express the concept of an UTI equilibrium as follows:

**Definition 6.3 (UTI Equilibrium)** A protocol $\alpha$ is an UTI equilibrium if and only if:

$$\forall J \in 2^{Agt^*}, \forall \beta \in \Delta, \text{EV}_J(\beta^J \cdot \alpha^{-J}) \leq \text{EV}_J(\alpha^J \cdot \alpha^{-J})$$

In order to illustrate the above UTI structure, let us consider the game presented in Section 6.1. The corresponding structure $UTI = \langle Agt, \{S_i | i \in Agt\}, \{U_J | J \in 2^{Agt^*}\}, \{\Omega_i | i \in Agt\} \rangle$ can therefore be defined as follows:

- $Agt = \{a, b\}$ where $a$ and $b$ respectively stand for Alice and Bob;
- $S_a = S_b = \{C, D\}$; $S_{\{a,b\}} = \{(D, D), (D, C), (C, D), (C, C)\}$;
- The individual payoff functions are defined according to Figure 6.1 (where $0 < y < x$):
  - $U_a(C, C) = U_b(C, C) = x$;
  - $U_a(D, D) = U_b(D, D) = y$;
  - $U_a(D, C) = U_b(C, D) = z$;
  - $U_a(C, D) = U_b(D, C) = 0$;
- The collective payoff function may then be freely defined according to Rawls’ criterion of fairness\(^1\), e.g.,
  - $U_{\{a,b\}}(C, C) = x$;
  - $U_{\{a,b\}}(D, D) = y$;
  - $U_{\{a,b\}}(C, D) = U_{\{a,b\}}(D, C) = 0$;
- for every $i \in \{a, b\}$, $\Omega_i$ is defined as follows:
  - $\Omega_i(\{a\}) = 1 - \omega_i$
  - $\Omega_i(\{b\}) = \omega_i$
  - $\Omega_i(\{i\}) = \omega_i$

where $\omega_i$ and $1 - \omega_i$ respectively characterize the probability that player $i$ reasons in the we-mode / I-mode.

In the case of the preceding game, let us consider the protocol $\alpha$ such that $\alpha(a, \{a\}) = \alpha(b, \{b\}) = \alpha(a, \{a, b\}) = \alpha(b, \{a, b\}) = C$. Protocol $\alpha$ simply specifies that both players choose strategy $C$ independently of reasoning in the I-mode or in the we-mode. Similarly, let us consider the alternative protocol $\beta$ such that

\(^1\)One could similarly use the classical utilitarian criterion as the collective payoff function.
\[ \beta(a, \{a\}) = \beta(b, \{b\}) = D \quad \text{and} \quad \beta(a, \{a, b\}) = \beta(b, \{a, b\}) = C. \]

In this case, protocol \( \beta \) specifies that both players choose strategy \( D \) when reasoning in the \textit{I-mode} and choose strategy \( C \) when reasoning in the \textit{we-mode}.

Table 6.2 then illustrates the conditions under which either of the above protocols (\( \alpha \) or \( \beta \)) is the unique \textit{UTI} equilibrium in the above \textit{UTI} structure, depending on the value of payoff \( z \) and the group identification parameters \( \omega_a \) and \( \omega_b \).

<table>
<thead>
<tr>
<th>collective payoff functions</th>
<th>Conditions on ( z )</th>
<th>Conditions on ( \omega_a ) and ( \omega_b )</th>
<th>Unique \textit{UTI} equilibria</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{maximin}</td>
<td>( 0 \leq z &lt; x )</td>
<td>( \frac{\omega_a \omega_b}{\omega_a + \omega_b} &gt; \frac{y}{x+y} )</td>
<td>( \alpha )</td>
</tr>
<tr>
<td></td>
<td>( x \leq z )</td>
<td>( \frac{\omega_a \omega_b}{\omega_a + \omega_b} &gt; \frac{y}{x+y} )</td>
<td>( \beta )</td>
</tr>
<tr>
<td>\textit{classical utilitarianism}</td>
<td>( 0 \leq z &lt; y )</td>
<td>( \frac{\omega_a \omega_b}{\omega_a + \omega_b} &gt; \frac{2y-z}{2x+2y-2z} )</td>
<td>( \alpha )</td>
</tr>
<tr>
<td></td>
<td>( y \leq z &lt; x )</td>
<td>( \max(\omega_a, \omega_b) &gt; \frac{y}{x+y-z} )</td>
<td>( \alpha )</td>
</tr>
<tr>
<td></td>
<td>( z &lt; z \leq 2x )</td>
<td>( 0 \leq \omega_a, \omega_b \leq 1 )</td>
<td>( \beta )</td>
</tr>
<tr>
<td></td>
<td>( 2x &lt; z )</td>
<td>( \max(\omega_a, \omega_b) &lt; \frac{z/2-y}{z-x-y} )</td>
<td>( \beta )</td>
</tr>
</tbody>
</table>

Table 6.2: Conditions for protocols \( \alpha \) and \( \beta \) to be a unique \textit{UTI} equilibrium

The main observation that can be made from Table 6.2 is that, whenever \( 0 \leq z < x \) (e.g., in the Hi-Lo game or the Stag hunt game), for some sufficiently high probability \( \omega_a \) or \( \omega_b \) that either agent \( a \) or \( b \) identifies as a member of team \( \{a, b\} \), playing strategy \( C \) becomes the only optimal choice no matter whether \( a \) and \( b \) actually reason in the \textit{I-mode} or in the \textit{we-mode} (cf. protocol \( \alpha \)) and independently of the collective payoff function (\textit{maximin} or classical utilitarianism).

However, there exist some notable differences between using Rawls’ criterion of fairness (\textit{maximin}) or the principle of classical utilitarianism as the collective payoff function whenever \( x \leq z \) (e.g., in the Prisoner’s dilemma). In fact, for some sufficiently high probability of identifying with the group (i.e., \( \omega_a \) and \( \omega_b \) are sufficiently large), following Rawls’ criterion of fairness leads to \( \beta \) being the unique \textit{UTI} equilibrium whenever \( z \geq x \). On the other hand, when using the classical utilitarianism criterion, the determination of a \textit{UTI} equilibrium is more complex as it relies on some stronger restriction on the value of \( z \). Indeed, protocol \( \beta \) is the unique optimal solution only if mutual cooperation is the best outcome for the group (i.e., when \( x < z < 2x \)), no matter the probability of each player identifying with the group (if one identifies with the group, then one will always play \( C \), else one will always play \( D \)).

\(^1\)Note that, whenever \( z = x \), then there exists no unique \textit{UTI} equilibrium: both protocols \( \alpha \) and \( \beta \) are \textit{UTI} equilibria.
outcome for the group (i.e., when $z > 2x$), if both players sufficiently identify with the group and $\omega_a = \omega_b$, then there exists no unique UTI equilibrium ($(C, D)$ and $(D, C)$ become equally good for both players).

Let us now give a more precise interpretation of a team reasoning structure in terms of a classical strategic form game.

**Definition 6.4 (Induced Strategic Game)** Given an unreliable team interaction structure $UTI = \langle Agt, \{S_i| i \in Agt\}, \{U_J| J \in 2^{Agts}\}, \{\Omega_i|i \in Agt\} \rangle$, the corresponding induced strategic game is a tuple $G^{uti} = \langle Agt', \{S'_J| J \in Agt'\}, \{U'_J|J \in Agt'\} \rangle$ where:

- $Agt' = 2^{Agts}$;
- for each $J \in Agt'$, $S'_J = S_J$;
- for each $J \in Agt'$, $U'_J$ is the utility function on $S'$ such that $U'_J(s) = EV_J(\alpha)$ where $\alpha \in \Delta$ is the protocol such that, for every $H \in Agt'$, $\alpha^H = s_H$.

According to Definition 6.4, every UTI structure induce a strategic form game with a player for every coalition in the original game. For example, in the above UTI structure for the game from Section 6.1, the induced strategic game would include three players, i.e., Alice, Bob, and the team whose members are Alice and Bob.

As already shown in Bacharach [1999], it is straightforward to demonstrate the following proposition.

**Proposition 6.2.1.1** Given an unreliable team interaction structure $UTI = \langle Agt, \{S_i| i \in Agt\}, \{U_J| J \in 2^{Agts}\}, \{\Omega_i|i \in Agt\} \rangle$, the protocol $\alpha \in \Delta$ is an UTI equilibrium if and only if the strategy profile $s$ such that $s_J = \alpha^J$ for all $J \in 2^{Agts}$ is a Nash equilibrium in the strategic game $G^{uti}$ induced by UTI.

### 6.2.2 Limitations

Although Bacharach’s theory of team reasoning clearly allows to model complex situations in which different competing coalitions may coexist, it has however some limitations that we want to discuss here.

As pointed out in Section 6.2, Bacharach’s theory relies on the assumption that every agent identifies with a unique team at a given time. This is a strong assumption, as it prevents from modeling situations in which an agent plays as a member of more than one group. To illustrate this, consider a scenario where one...
faces the dilemma between cooperating with a close friend, and cooperating with a family member (assuming the friend and the family member cannot cooperate with each other). In this case, Bacharach’s theory predicts that the individual will choose over these two options, even though a more egalitarian solution might exist that would satisfy equally all players (some concrete examples are discussed in Sections 6.3.4 and Appendix C.2). This limitation is therefore particularly relevant when modeling more flexible and heterogenous multiagent systems in which different coalitions might be formed whose intersections are non-empty.

Another problem of the theory of team reasoning concerns the exogenous probabilistic distributions $\Omega_i$ in the definition of an UTI structure. In fact, it is not completely clear how they should be interpreted. While such probabilities may depend on some intrinsic features of the game such as the payoff structure, they may also be determined by some pre-existent social relationships between the players. Bacharach’s theory however remains vague regarding this issue.

Apart from the previous conceptual restrictions and ambiguities, the theory of team reasoning has a technical limitation which lies in the complexity of the problem of computing equilibria in the context of teamwork by using Nash equilibrium (see Proposition 6.2.1.1). In fact, as Definition 6.4 indicates, given a structure $UTI = \langle \text{Agt}, \{S_i|i \in \text{Agt}\}, \{U'_J|J \in 2^{\text{Agt}}\}, \{\Omega_i|i \in \text{Agt}\} \rangle$, the size of the set of agent $\text{Agt}'$ in the game $G^{uti}$ induced by $UTI$ is exponential in $\text{Agt}$, in particular we have $|\text{Agt}'| = 2^{|\text{Agt}|} - 1$. In other words, if one wants to use Nash equilibrium in order to compute $UTI$ equilibria, he has to increase exponentially the size of the structure of interaction.

In the next section we present a generalization of our model of social ties from Chapter 4 (see Section 4.2), which provides a simpler approach to group reasoning and collective decision making. The interesting aspect of this model is that it allows to compute equilibrium solutions in the context of collaborative activity by using the concept of Nash equilibrium and without increasing the size of the structure of interaction.

### 6.3 A theory of social ties

In this section, we introduce a new model that characterizes well the agents’ behavior in the presence of social ties. Such a model, which extends that from Section 4.2 in Chapter 4, aims at explaining social dilemmas as in the game from Section 6.1 by simply following the main idea on which theories of social preferences are based (see, e.g., Charness and Rabin [2002]; Fehr and Schmidt [2006]), that is: (1) performing some utility transformation in a given game, and then (2) of applying classical solution concepts from game theory (e.g., Nash equilibrium) in the transformed game in order to find equilibria. Indeed, similarly to theories of social...
preferences, our starting assumption is that the existence of a social tie between two individuals may have an impact on their utilities, in the sense that the utility that an agent attaches to a given outcome may be affected by his social ties with other agents. In the next section, we explain in detail how social ties may affect an agent’s utility function.

### 6.3.1 Definitions

Let us introduce our social ties game, which extends the strategic game structure presented in Definition 6.1 (see Section 6.1).

**Definition 6.5 (Social Ties Game)** A social ties game is a tuple \( ST = \langle \text{Agt}, \{S_i \mid i \in \text{Agt}\}, \{U_J \mid J \in 2^{\text{Agt}}\}, \{k_i \mid i \in \text{Agt}\} \rangle \) where:

- \( \langle \text{Agt}, \{S_i \mid i \in \text{Agt}\}, \{U_J \mid J \in 2^{\text{Agt}}\} \rangle \) is a strategic game with group utility according to Definition 6.1;
- every \( k_i \) is a total function \( k_i : \text{Group}(i) \rightarrow [0, 1] \), such that:
  - **C1** for every \( i \in \text{Agt} \), \( \sum_{J \in \text{Group}(i)} k_i(J) = 1 \)
  - **C2** for all \( i, j \in J \), \( k_i(J) = k_j(J) \)

Every parameter \( k_i(J) \) in Definition 6.5 should be seen as a measure of the social tie between agent \( i \) and group \( J \) given the current game context. In particular, \( k_i(J) \) measures the degree with which agent \( i \) identifies with group \( J \). Setting \( k_i(J) \) to 0 corresponds to a non-existing tie between agent \( i \) and group \( J \) (i.e., \( i \) does not identify with \( J \)), whereas setting \( k_i(J) \) to 1 means that agent \( i \) is strongly tied with group \( J \) (i.e., \( i \) strongly identifies with \( J \)). Moreover, note that \( k_i(\{i\}) \) stands for agent \( i \)'s measure of individualism.

According to Constraint **C1**, an agent’s identifications with different groups sum up to one, i.e., an agent can neither fully identify with different groups, nor identify with no group at all (in the most extreme case, the agent \( i \) is maximally individualistic in the sense that \( k_i(\{i\}) = 1 \)). Through Constraint **C2**, we assume that a social tie is restricted to be bilateral, which can be interpreted as follows: the degree of the social tie with some group \( J \) is the same for every member of \( J \).

The following definition introduces the notion of Social Ties utility function, i.e., how an agent’s utility is affected by his social ties.

**Definition 6.6 (Social Ties Utility)** For every strategy profile \( s \in S \), the Social Ties utility function of player \( i \) is given by:

\[
U_i^{ST}(s) = \sum_{J \subseteq \text{Agt}\setminus\{i\}} k_i(J \cup \{i\}) \cdot \max_{s'_J \in S_J} U_{J \cup \{i\}}(s_{-J}, s'_J)
\]

157
In Definition 6.6, every $s_{-J} \in S_{-J}$ denotes a joint strategy for the coalition $\text{Agt}\setminus J$ (i.e., $S_{-J} = S_{\text{Agt}\setminus J}$), and $U_J(s)$ stands for group $J$’s utility function, as it is described in Definition 6.1.

The general idea of our model, which is formally expressed by the preceding Social Ties utility function $U^i_{ST}(s)$, is that, in the presence of a strong social tie between an individual $i$ and some group $J \in \text{Group}(i)$, agent $i$ will be motivated to maximize the benefit of group $J$ represented by collective utility $U_J$, assuming that agents in $J$ are also motivated to maximize the benefit of group $J$. Note that, in this case, agent $i$ does not face a full strategic problem anymore. Indeed, the utility of the strategy profile $s$ for agent $i$ becomes independent of the strategies of members of group $H = J \setminus \{i\}$ (i.e., $s_H$). Therefore, agent $i$ only needs to reason strategically regarding the choices of every player outside of $J$, and choose the action from the strategy profile which maximizes group $J$’s utility.

As in the previous section, let us now interpret the above social ties game in terms of a classical strategic form game.

**Definition 6.7 (Induced Strategic Game)** Given a social ties game $ST = \langle \text{Agt}, \{S_i|i \in \text{Agt}\}, \{U_J|J \in 2^{\text{Agt}^*}\}, \{k_i|i \in \text{Agt}\}\rangle$, the corresponding induced strategic game is a tuple $G^ST = \langle \text{Agt}, \{S_i|i \in \text{Agt}\}, \{U'_i|i \in \text{Agt}\}\rangle$ where, for all $i \in \text{Agt}$ and all $s \in S$:

$$U'_i(s) = U^i_{ST}(s)$$

As a concrete example of such a social ties game, let us again consider the type of interactions illustrated in Figure 6.1 from Section 6.1. Let us first restate the Social Ties utility function from Definition 6.6, which can be simplified as follows when applied to any two player game. For any $i, j \in \text{Agt}$ such that $i \neq j$, for every $s \in S$:

$$U^i_{ST}(s) = (1 - k_{ij}) \cdot U_i(s) + k_{ij} \cdot \max_{s'_j \in S_j} U_{\{i,j\}}(s_i, s'_j)$$

where $k_{ij}$ stands for $k_i(\{i, j\})$.

Note that this simplification simply corresponds to the previous utility function introduced in Section 4.2 from Chapter 4. Starting from the matrix payoff from Figure 6.1, Figure 6.2 represents the corresponding transformed utilities for each player based on Rawls’ maximin principle as the group utility function, and where an extremely strong social tie exists between Alice and Bob (i.e., $k_{ab} = k_{ba} = 1$).

In this case, it is easy to show, through iterated elimination of strictly dominated strategies, that the only Nash equilibrium resulting from this transformation is $(C, C)$. One should note that each player’s choice becomes independent of the opponent’s choice. In other words, the initially strategic problem becomes a classical problem of individual decision making that each player has to solve. More
generally, Table 6.3 depicts the predictions that can be made in this game depending on the value of \( z \) and social tie parameters \( k_{ab} \) and \( k_{ba} \).

<table>
<thead>
<tr>
<th>collective payoff functions</th>
<th>Conditions on ( z )</th>
<th>Conditions on ( k = k_{ab} = k_{ba} )</th>
<th>Unique Nash equilibria</th>
</tr>
</thead>
<tbody>
<tr>
<td>maximin</td>
<td>( 0 \leq z \leq x+y )</td>
<td>( k &gt; \frac{y}{x} )</td>
<td>( (C,C) )</td>
</tr>
<tr>
<td></td>
<td>( x+y \leq z )</td>
<td>( k &gt; \frac{z-x}{z-y} )</td>
<td></td>
</tr>
<tr>
<td>classical utilitarianism</td>
<td>( 0 \leq z \leq 2y )</td>
<td>( k &gt; \frac{y}{2x+y-z} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( 2y \leq z &lt; 2x )</td>
<td>( k &gt; \frac{y}{2x+y-z} ) and ( k &gt; \frac{z-x}{x} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( 2x \leq z )</td>
<td>( 0 \leq k &lt; 1 )</td>
<td>( (D,D) )</td>
</tr>
</tbody>
</table>

Table 6.3: Notable predictions in the strategic game induced by \( ST \)

The main observation one can make from Table 6.3 is that, as long as there exists a unique best outcome for the group in the above game (i.e., whenever \( 0 \leq z < 2x \)), strategy \( C \) strictly dominates \( D \) in the context of some sufficiently strong social tie between both players. However, one should note that, as for team reasoning (see Table 6.2 in Section 6.2), some differences appear between the predictions made by the two types of collective payoff functions. The most notable distinction concerns the particular case where there exist conflicting collective goals (i.e., when \( z \geq 2x \), \( (C,D) \) and \( (D,C) \) become equally best outcomes for the group). In this case, under the assumption of the maximin principle, strong social ties still lead to mutual cooperation. On the other hand, when using the classical utilitarianism criterion, strong social ties allow for the unique most optimal outcome to be reached through mutual defection (as if the players were purely self-regarding).
6.3.2 Analysis

The following theorem demonstrates the ability of social ties to converge towards playing a Nash equilibrium solution.

**Theorem 6.1** For any strategic game with group utility $G = \langle \text{Agt}, \{S_i | i \in \text{Agt}\}, \{U_J | J \in 2^{\text{Agt}}\} \rangle$, there exists a social ties game $ST = \langle G, \{k_i | i \in \text{Agt}\} \rangle$ whose induced strategic game has a Nash equilibrium in pure strategies.

One can note, through the proof of Theorem 6.1 in the Appendix, that there exist games with group utility $G$ that do not yield a unique Nash equilibrium in the strategic game induced by any social ties game $ST = \langle G, \{k_i | i \in \text{Agt}\} \rangle$. In this case, as several distinct equilibria are unable to make any clear prediction, one should note that exploiting the group utility as done by such a $ST$ game therefore becomes irrelevant. However, notice that, for any game with group utility $G = \langle \text{Agt}, \{S_i | i \in \text{Agt}\}, \{U_J | J \in 2^{\text{Agt}}\} \rangle$ such that $\text{argmax}_{s' \in S} U_{\text{Agt}}(s')$ is a singleton (i.e., the group $\text{Agt}$ does not have conflicting goals), there exists a social ties game $ST = \langle G, \{k_i | i \in \text{Agt}\} \rangle$ such that $k_i(\text{Agt}) = 1$ for every $i \in \text{Agt}$, and the game induced by $ST$ has a unique pure strategy Nash equilibrium.

6.3.3 Relationship with team reasoning

Let us now provide a comparative analysis of the social ties game with the previous theory of team reasoning (see Section 6.2). For this purpose, we first present the main advantage of our model of social ties over Bacharach’s theory regarding the problem of gradual group identification. We then restrict the comparative analysis of both models in the particular context of two-player games before extending it to any $n$-player game such that $n \geq 2$.

6.3.3.1 Gradual group identification

As already mentioned in Chapter 4 (see Section 4.8), Bacharach’s theory of team reasoning and our model of social ties have an important difference that is worth recalling here. While Bacharach’s theory assumes that an agent in a given strategic setting reasons either in the $I$-mode or in the $we$-mode, our theory of social ties assumes that an agent can be partially tied with a given group or team. In other words, differently from Bacharach’s theory of team reasoning, our theory of social ties allows to model a notion of partial identification with a group. As a result of this difference, it appears that both game structures can disagree about the predicted outcome.

---

1This remark also applies to the theory of team reasoning from the previous section.
In order to illustrate this argument, let us formalize the simple two player dictator game introduced in Section 4.8 from Chapter 4, which can be defined as the strategic game with group utility \( G = \langle \{i, j\}, \{S_i, S_j\}, \{U_i, U_j, U_{\{i, j\}}\} \rangle \). According to \( G \), player \( i \) can choose between three options, i.e., \( S_i = \{A, B, C\} \), and player \( j \) is not facing any decision problem, i.e., \( S_j = \{D\} \). In such a scenario, each player’s payoff is determined uniquely from these strategies according to Table 6.4 (i.e., player \( j \) has no control over the outcome, which is uniquely determined by player \( i \)). Note that the collective payoff function can then be computed both in terms of global efficiency (i.e., sum of individual payoffs: \( U_{\{i, j\}} = U_i(s) + U_j(s) \)) and in terms of the Rawlsian criterion of fairness (i.e., minimum of individual payoffs: \( U_{\{i, j\}} = \min\{U_i(s), U_j(s)\} \)).

<table>
<thead>
<tr>
<th>Player ( i )'s option ( s_i \in S_i )</th>
<th>( U_i )</th>
<th>( U_j )</th>
<th>( U_{{i, j}} = U_i(s) + U_j(s) )</th>
<th>( U_{{i, j}} = \min{U_i(s), U_j(s)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>( B )</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>( C )</td>
<td>7</td>
<td>4</td>
<td>11</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 6.4: Simple dictator game

Applying team reasoning to this particular situation leads to the following predictions: in any \( UTI \) structure \( UTI = \langle G, \{\Omega_i, \Omega_j\} \rangle \), player \( i \) will play \( A \) if reasoning in \( I\)-mode (i.e., player \( i \) is then self-regarding), and player \( i \) will play \( B \) if reasoning in \( we\)-mode (i.e., player \( i \) then identifies with the group). As a consequence, according to Bacharach’s theory of team reasoning (and independently of whether the collective utility function is computed using the \( \text{sum} \) or the \( \text{min} \)), player \( i \) will never happen to choose \( C \): there indeed exists no \( UTI \) structure with a probability distribution \( \Omega_i \) of identifying with the group (0 ≤ \( \Omega_i(\{i, j\}) \) ≤ 1) that can specify this outcome to occur (in either the \( I\)-mode, or in the \( we\)-mode). Note that in any such \( UTI \) structure, player \( j \)’s mode of reasoning (through function \( \Omega_j \)) is irrelevant to the determination of an equilibrium solution.

On the other hand, considering the same game through our model of social ties leads to a different interpretation: in some social ties game \( ST = \langle G, \{k_i, k_j\} \rangle \), player \( i \) will select \( A \) if both players are extremely close to each other (e.g., \( k_i(\{i, j\}) = 1 \)), and player \( i \) will select \( B \) if they instead are perfect strangers (e.g., \( k_i(\{i\}) = 1 \)). However, if both players are neither best friends nor perfect strangers but, say, simple acquaintances (e.g., \( k_i(\{i, j\}) = k_i(\{i\}) = 0.5 \)), then player \( i \) will choose \( C \), as a compromise between being self-regarding and group-regarding. Indeed, one can easily prove that if \( k_i(\{i, j\}) = k_i(\{i\}) = 0.5 \), then action \( C \) is the unique Nash equilibrium in the transformed game where utilities
are computed using Definition 6.6, both when the collective utility function is computed using the *sum* and when it is computed using the *min*. In other words, our model of social ties predicts that if player $i$ *partially* identifies as a member of group $\{i,j\}$, then he will choose option $C$. However, although this seems a reasonable conclusion, it appears to be inconsistent with Bacharach’s theory of team reasoning, as shown above. Note that a more detailed analysis of the game from Table 6.4 is provided in Appendix C.2.

Furthermore, as already mentioned in Section 6.2.2, such a difference between both game structures can be emphasized by considering more complex multiagent systems that involve the formation of sub-coalitions whose intersections are non-empty. A concrete illustration of such complex social interactions is provided in Section 6.3.4.

### 6.3.3.2 Similarities in any two-player game

In the context of two-player games, let us first specify the various similarities that can emerge when considering subclasses of the above models of team reasoning and social ties. In order to do so, we then consider some binary interpretation of those game structures, which can be defined as in Definitions 6.8 and 6.9.

**Definition 6.8 (Binary UTI structure)** A binary unreliable team interaction BUTI is a structure $UTI = \langle \text{Agt}, \{S_i| i \in \text{Agt}\}, \{U_{J}| J \in 2^{\text{Agt}}\}, \{\Omega_i|i \in \text{Agt}\}\rangle$ where there exists $g \in \text{Groups}$ such that:

- for every $i \in \text{Agt}$, $\Omega_i(g_i) = 1$;
- for all $i, j \in \text{Agt}$, we have that either $g_i = g_j$, or $g_i \cap g_j = \emptyset$.

According to the BUTI structure, there is no uncertainty about which group each agent identifies with. In such a structure, the concept of a protocol can be reduced to a simple strategy profile, which therefore allows the comparison with some ST game. Moreover, note that in this case, an agent identifies with a group if and only if other members of that group also identify with it.

**Definition 6.9 (Binary ST Game)** A binary social ties game BST is a game $ST = \langle \text{Agt}, \{S_i| i \in \text{Agt}\}, \{U_{J}| J \in 2^{\text{Agt}}\}, \{k_i|i \in \text{Agt}\}\rangle$ where:

- for every $i \in \text{Agt}$ and every $J \in \text{Group}(i)$, $k_i(J) \in \{0,1\}$.

Similarly, the BST game represents an extreme interpretation of the ST game where every agent can only identify with a unique group.

Thus performing a detailed analysis of such binary two-player games can reveal their similarities, as shown through Theorem 6.2.
**Theorem 6.2** Given a strategic game with group utility \( G = \langle \{i, j\}, \{S_i, S_j\}, \{U_i, U_j, U_{\{i,j\}}\} \rangle \) where \( \text{argmax}_{s \in \mathcal{S}} U_{\{i,j\}}(s) \) is a singleton, let \( \text{BST} = \langle G, \{k_i, k_j\} \rangle \) be a binary social ties game, and \( \text{BUTI} = \langle G, \{\Omega_i, \Omega_j\} \rangle \) a binary unreliable team interaction structure such that \( k_i(\{i, j\}) = k_j(\{i, j\}) = \Omega_i(\{i, j\}) = \Omega_j(\{i, j\}) \).

If \( s \in \mathcal{S} \) and \( \alpha \in \Delta \) are such that \( \alpha^{\{i,j\}} = (\alpha^{(i)}, \alpha^{(j)}) = s \), then \( s \) is a Nash equilibrium in the game induced by \( \text{BST} \) if and only if \( \alpha \) is an UTI equilibrium in \( \text{BUTI} \).

Theorem 6.2 therefore indicates that binary versions of both game structures can make the same predictions regarding the agents’ behavior in any two-player game.

Moreover, one should note that the type of game that is considered in Theorem 6.2 is restrictive regarding what determines the best outcome of the group. In fact, another common property that both of the above models share is their reliance on the concept of collective goals. In an UTI structure as well as in a ST game, the formation of a group implies that all members of this group seek to satisfy the group’s objective, that is, they aim at reaching the highest possible payoff for the group. However, it appears that, depending on the type of interactive situation being considered, such a collective goal may not be clearly determined. As an example, one may consider the two-player game in Figure 6.3, which corresponds to the well-known Battle of the Sexes game.

\[
\begin{array}{c|cc}
 & D & R \\
\hline
U & (10, 5) & (0, 0) \\
D & (0, 0) & (5, 10) \\
\end{array}
\]

**Figure 6.3: Battle of the Sexes**

In the context of this game, it is straightforward to show that any UTI structure and any ST game will always be indecisive, no matter the type of group identification involved. Such an observation is obviously justified by the fact that the group made of the two players does not have a unique goal (i.e., both \((U, D)\) and \((D, R)\) are equally good for the group according to either classical utilitarianism or the maximin principle). As a consequence, our theory of social ties and Bacharach’s theory of team reasoning are simply unable to make any prediction in such a game (note that Theorem 6.2 does not apply to games as in Figure 6.3).
6.3.3.3 Further comparison in any n-player game

Let us now extend the previous analysis with considering more complex interactive situations that involve more than two players.

We have demonstrated in Section 6.3.3.2 that both models UTI and ST face the same limitation whenever some collective goals are conflicting in two-player games. We now intend to show that, although this remark remains true, conflicting collective goals are interpreted differently by the two theories.

As a means to emphasize this difference between our model of social ties and bacharach’s theory of team reasoning, we define a restricted class of strategic games with non-conflicting collective goals according to Definition 6.10.

Definition 6.10 (Game with Non-conflicting Collective Goals) A game with non-conflicting collective goals is a game with group utility \( G = \langle \text{Agt}, \{S_i|i \in \text{Agt}\}, \{U_i|i \in \text{Agt}\}\rangle \) as defined in Definition 6.1, such that:

\[ C3 \] for every \( J \in 2^{\text{Agt}} \) s.t. \( |J| > 1 \) and every \( s_{-J} \in S_{-J} \), \( \arg\max_{s' \in S_J} U_J(s'_J, s_{-J}) \) is a singleton.

In Definition 6.10, Constraint C3 simply ensures that no team can have multiple conflicting goals at once. Note that this constraint indeed rules out interactive situations such as the game in Figure 6.3.

Thus performing a detailed analysis of a binary n-player game can reveal their equivalence under the previous constraint C3, as shown through Theorem 6.3.

Theorem 6.3 Given a strategic game with group utility \( G = \langle \text{Agt}, \{S_i|i \in \text{Agt}\}, \{U_J|J \in 2^{\text{Agt}}\}\rangle \) that satisfies Constraint C3 from Definition 6.10, let BST = \( \langle G, \{k_i|i \in \text{Agt}\}\rangle \) be a binary social ties game, and BUTI = \( \langle G, \{\Omega_i|i \in \text{Agt}\}\rangle \) a binary unreliable team interaction structure such that, for every \( i \in \text{Agt} \) and every \( J \in \text{Group}(i) \), \( k_i(J) = \Omega_i(J) \).

If \( s \in S \) and \( \alpha \in \Delta \) are such that, for every \( J \in 2^{\text{Agt}} \), \( \alpha^J = s_J \), then \( s \) is a Nash equilibrium in the game induced by BST if and only if \( \alpha \) is an UTI equilibrium in BUTI.

Theorem 6.3 therefore generalizes Theorem 6.2 from Section 6.3.3.2 and shows that binary versions of both game structures can make the same predictions regarding the agents’ behavior in any n-player game that satisfies Constraint C3.

However, the need for an additional constraint on the original game structure \( G \) in Theorem 6.3 (i.e., Constraint C3) points out to another important conceptual difference concerning the type of transformation each model is based upon.
Indeed, Bacharach’s concept of team reasoning relies on what he calls agency transformation, which consists in conceiving the situation not as a decision making problem for individual agents, but as a decision making problem for the group as an agent. Alternatively, our concept of group identification relies on the idea that social ties influence players’ utilities without interfering with their type of reasoning (i.e., they may then apply the classical principle of rationality to such transformed utilities).

As a means to illustrate the different predictions that can be made by our theory of social ties and by Bacharach’s theory of team reasoning whenever Constraint C3 from Definition 6.10 is removed, let us consider the three player game depicted in Table 6.5.

<table>
<thead>
<tr>
<th>Actions</th>
<th>Utilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Player 1</td>
</tr>
<tr>
<td></td>
<td>Player 1</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

Table 6.5: A three-player coordination game

One can indeed observe from the game in Table 6.5 that the group made of player 1 and player 2 appears to have conflicting goals if player 3 chooses A, that is, both (A,A,A) and (B,B,A) are equally best outcomes for group {1,2}. In this case, considering an UTI structure depicting a strong group identification with group {1,2} (i.e., \(\Omega_1(\{1,2\}) = \Omega_2(\{1,2\}) = 1\)) while player 3 is strictly individualistic (i.e., \(\Omega_3(\{3\}) = 1\)), Bacharach’s theory predicts that player 1 and player 2 should both select option A. In fact, if player 3 chooses A, then both player 1 and player 2 are indecisive between playing either (A,A) or (B,B) (both strategies then yield the same collective payoff). However, in the case they play (B,B), then player 3 would be better off selecting B. Consequently, the only equilibrium solution that can be found in this game is for the team {1,2} to perform (A,A).

On the other hand, if considering the same scenario through a social ties game (i.e., \(k_1(\{1,2\}) = k_2(\{1,2\}) = 1\) and \(\Omega_3(\{3\}) = 1\)), a different interpretation arises. If player 3 chooses B, then the only optimal strategy for both player 1 and player 2 is to select A. However, if player 3 chooses A, then, as a result of being uncertain
about what is the right thing to do for the group, both options A and B become equally good for player 1 and player 2. Consequently, the group \{1, 2\} may come to miscoordinate and play either (A, B) or (B, A): for example, player 1 may select A, assuming that player 2 will aim at maximizing the benefit of the group by also selecting A, while player 2 actually selects B, similarly assuming that player 1 will aim at maximizing the benefit of the group by also selecting B. As a result, our theory of social ties does not allow to predict a unique solution in this particular configuration of social ties in the game from Table 6.5 (See Appendix C.4 for more details).

More generally, this scenario illustrates the fact that, unlike Bacharach’s theory of team reasoning, our theory of social ties does not rely on a strong notion of unity between individuals. In fact, it assumes that individuals do not act as if they were a single agent (which is implied by Bacharach’s concept of agency transformation), but instead they act as separated entities that simply share a common goal.

Moreover, another consequence of the different types of transformation each model relies on concerns the complexity for computing an equilibrium solution. In fact, unlike for Bacharach’s UTI structures, ST games consider the standard concept of a strategy profile, as in traditional game theory (UTI structures instead consider protocols, as depicted in Section 6.2.1). As shown through Definition 6.7, a consequence is that the strategic game \(G^{ST}\) induced by any social ties game \(ST\) does not increase the complexity of computing the equilibrium solution, which is indeed an important difference with Bacharach’s UTI structure: given a game with group utility \(G\), the game \(G^{ST}\) induced by any \(ST = \langle G, \{k_i | i \in Agt\}\rangle\) only transforms the utility function of the original game \(G\) (the number of agents remains unchanged here). Finding a Nash equilibrium in the game \(G^{ST}\) therefore appears to be mathematically simpler than finding such a solution in the game \(G^{uti}\) induced by any \(UTI = \langle G, \{\Omega_i | i \in Agt\}\rangle\).

6.3.4 Illustration: the Three Musketeers game

Through this section, we present a concrete scenario that allows to illustrate the formation of different coalitions whose intersections are non-empty, and its effect on the each individual’s strategic behavior.

Let us consider a situation involving three individuals named Athos, Porthos, and Aramis after Alexandre Dumas’s three Musketeers (from his famous historical novel titled “Les Trois Mousquetaires”). In this fictitious scenario, the three Musketeers have been arrested by Cardinal Richelieu who suspects them to have killed one of his most precious officers on the preceding evening. Richelieu reveals to all Musketeers that he was provided with some evidence proving that:

- Athos is not directly responsible for the crime (i.e., the guilty individual is
either Porthos or Aramis).

- Athos spent the preceding evening at another innocent Musketeer’s house.
- Porthos and Aramis do not know the exact location of each other’s house.

As an attempt to reveal who, among Porthos and Aramis, is guilty, he therefore proposes to interrogate all Musketeers individually about their respective location at the time of the crime: an individual is then considered guilty if he appears to be the only one having spent the evening alone. In this case, he indicates to all Musketeers that, if a guilty person is revealed, he will then be condemned to a death sentence while the others will be freed immediately. Furthermore, Richelieu also specifies that any detected lie will be severely punished by all three individuals thereby facing the same death sentence (i.e., if at least two of the Musketeers’ statements are inconsistent). Richelieu further indicates that if Athos states that he spent the evening at his own house (which is suspected to be false) and a guilty person is revealed, then he will still be freed but will lose his affiliation with the Musketeers of the Guard while the other remaining innocent musketeer will be promoted as the new captain of the Musketeers. On the other hand, in the particular case where all three individuals similarly state that they spent the evening together at Athos’s house (i.e., no lie is detected and no guilty person is revealed), then they will all be held equally responsible and will consequently have to equally serve some time in jail. In the meantime, the Cardinal admits that, if all individuals indicate having spent the evening alone at their own respective house, he will not have enough evidence to convict any of them on the principal charge. In this case, while all individuals will then be released, they will simply lose their affiliation with the Musketeers of the Guard.

A formal representation of this scenario is depicted in Table 6.6. Actions A, B, and C stand respectively for “spending the evening at Porthos’s house”, “spending the evening at Aramis’s house”, and “spending the evening at Athos’s house”. Note that spending the evening at an individual i’s home naturally implies the presence of i there. Any statement that contradicts this principle will therefore be considered as a lie by the Cardinal. We then assign, for each Musketeer, a payoff of 6 for an immediate release along with becoming the captain of the Musketeers of the Guard, a payoff of 5 for a simple immediate release (without any promotion), a payoff of 4 for an immediate release along with the loss of affiliation with the Musketeers of the Guard, a payoff of 3 for a prison sentence, and a payoff of 0 for a death sentence.

One should note that the game in Figure 6.6 may be considered as a combination of the simple games presented in Section 6.1. In fact, one can observe that the interaction between Athos and Porthos is similar to the Hi-Lo game. Similarly, this remark also applies to the type of interaction existing between Athos and
Aramis. Moreover, whenever Athos plays $C$, then the interaction between Porthos and Aramis simply corresponds to a Prisoner’s Dilemma.

All along the following cases, we assume that the collective group utility function is based on the Rawlsian criterion of fairness described in Section 6.1.

Let us first assume that the players are purely self-interested, that is, for every $i \in \{\text{Athos, Porthos, Aramis}\}$, $k_i(\{i\}) = 1$ in the corresponding social ties game. In this case, the game depicted in Table 6.6 yields the following three distinct Nash equilibria in pure strategy:

- $(A, A, \cdot)$: Athos and Porthos both play $A$ (Aramis’s action is irrelevant here);
- $(B, B, \cdot)$: Athos and Aramis both play $B$ (as before, Aramis’s action is also irrelevant here);
- $(C, C, C)$: Athos, Porthos, and Aramis all play $C$.

As a result of having multiple equilibria, no reasonable prediction can be made regarding each individual’s choice here.

However, according to Dumas’s original story, those Musketeers are considered to be inseparable friends who live by the motto “one for all, and all for one”. This interpretation can therefore imply that strong social ties exist between them (i.e., for every $i \in \{\text{Athos, Porthos, Aramis}\}$, $k_i(\{i\}) = 1$ in the social ties game) and therefore all players identify with the same unique team. In this particular case, the unique Nash equilibrium that results is $(C, A, B)$ where all Musketeers maximize the group’s preferences.

<table>
<thead>
<tr>
<th>Actions</th>
<th>Payoffs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athos</td>
<td>Porthos</td>
</tr>
<tr>
<td>$A$</td>
<td>$A$</td>
</tr>
<tr>
<td>$A$</td>
<td>$A$</td>
</tr>
<tr>
<td>$A$</td>
<td>$C$</td>
</tr>
<tr>
<td>$A$</td>
<td>$C$</td>
</tr>
<tr>
<td>$B$</td>
<td>$A$</td>
</tr>
<tr>
<td>$B$</td>
<td>$A$</td>
</tr>
<tr>
<td>$B$</td>
<td>$C$</td>
</tr>
<tr>
<td>$B$</td>
<td>$C$</td>
</tr>
<tr>
<td>$C$</td>
<td>$A$</td>
</tr>
<tr>
<td>$C$</td>
<td>$A$</td>
</tr>
<tr>
<td>$C$</td>
<td>$C$</td>
</tr>
<tr>
<td>$C$</td>
<td>$C$</td>
</tr>
</tbody>
</table>

Table 6.6: The Three Musketeers game
On the other hand, other fictional but realistic scenarios that differ from the previous situations could also be considered. For example, one may assume that Athos and Porthos are still very good friends, but both of them strongly dislike Aramis. Intuitively, this implies that Athos and Porthos make their own coalition that competes with Aramis (i.e., for every $i \in \{\text{Athos}, \text{Porthos}\}$, $k_i(\{\text{Athos}, \text{Porthos}\}) = 1$ and $k_{\text{Aramis}}(\{\text{Aramis}, i\}) = 0$ in the social ties game). As a result, the unique Nash equilibrium in the game induced by the corresponding social ties game then becomes $(A, A, \cdot)$, that is, Athos and Porthos should both select $A$, independently of Aramis’s action. Following the same way of reasoning, assuming that Porthos is isolated by the two others who form a coalition (i.e., for every $i \in \{\text{Athos}, \text{Aramis}\}$, $k_i(\{\text{Athos}, \text{Aramis}\}) = 1$ and $k_{\text{Porthos}}(\{\text{Porthos}, i\}) = 0$ in the social ties game), then the unique solution would similarly become $(B, B, \cdot)$.

Finally, let us consider the most illustrative situation in the above game that consists in Athos being extremely close to both Porthos and Aramis while the latter individuals are extremely weakly tied with each other. Intuitively, it is clear that nobody should then identify with the whole group because Porthos and Aramis are not willing to collaborate with each other (i.e., for every $i \in \{\text{Athos}, \text{Porthos}, \text{Aramis}\}$, $k_i(\{\text{Athos}, \text{Porthos}, \text{Aramis}\}) = 0$). Instead, such a scenario implies the existence of two different sub-coalitions whose intersection is non-empty: one coalition is made of Athos and Porthos whereas the other is made of Athos and Aramis (i.e., for every $i \in \{\text{Porthos}, \text{Aramis}\}$, $k_{\text{Athos}}(\{\text{Athos}, i\}) = 0.5$ and $k_i(\{\text{Porthos}, \text{Aramis}\}) = 0$ in the social ties game). In this case, our theory specifies that, while Athos will aim at equally satisfying both of these teams, Porthos and Aramis will value their own respective sub-coalition exactly as much as they value their self-interest (i.e., for every $i \in \{\text{Porthos}, \text{Aramis}\}$, $k_i(\{i\}) = 0.5$ in the social ties game). Yet in this case, in spite of the obvious lack of unity existing among the three Musketeers (i.e., no individual will identify with the largest group), our theory of social ties specifies that the unique equilibrium solution is for all individuals to select $C$, which illustrates another important difference in interpretation that exists between our social ties game and Bacharach’s UTI structure. In fact, it can be shown that there exists no UTI structure based on this game that uniquely specifies that Athos should play $C$ under the assumption that identification with the largest group is negligible (i.e., when all players do not identify with the largest group as much as they identify with another sub-coalition). More specifically, whenever Athos is torn between satisfying group $\{\text{Athos}, \text{Porthos}\}$ and satisfying group $\{\text{Athos}, \text{Aramis}\}$, then Bacharach’s theory of team reasoning predicts that he will play either $A$ or $B$, which is clearly counter-intuitive. A more detailed mathematical analysis underlying this distinction can be found in Appendix C.5.

As a main result, the analysis of this game simply illustrates the way in which
social ties can be used as some equilibrium selection device: indeed, we have shown that, in this particular situation, a different equilibrium solution is predicted for each type of social ties.

6.4 A theory of empathetic preferences

In this section, we discuss another alternative theory that supports the intuition behind our model of social ties from the previous section: Binmore’s theory of empathetic preferences (Binmore [1994, 1998, 2005]).

According to such a theory, the concept of empathizing with some individual(s) simply corresponds to making a decision as if being in the so-called “original position”, as introduced in Rawls’s theory of justice (Rawls [1971]). The original position is an ideal situation in which agents have to make a collective agreement about the fundamental principles of justice defining the society. The main distinguishing feature of this original position is the so-called “veil of ignorance”, behind which every agent is assumed to be deprived of all knowledge about his personal identity and characteristics so that he imagines himself to possibly be in any player’s position. The veil of ignorance is conceived by Rawls as a device which insures impartiality of judgment. In fact, if the agents do not know who they are and under which circumstances they will be acting in the future, they will be more prone to choose principles of justice supporting the entire society rather than those supporting a single agent (or a minority of agents).

However, in order to use this original position, Binmore argues that an agent must be equipped with some empathetic preferences, which consist in combining his actual own preferences with his preferences when imagining himself to be in the other agents’ positions. In other words, an individual must be able to evaluate the options available to him when identifying with any player (including himself). As an example of such empathetic preferences, Alice may have to compare eating an apple while being herself with eating an orange while being Bob. In this case, Binmore points out that, when projecting herself to be in Bob’s position, Alice must not consider her own preferences, she must instead imagine herself while having Bob’s preferences: if Bob prefers to eat an apple rather than an orange, then Alice should share this preference when putting herself in Bob’s position, even though she might herself prefer to eat an orange rather than an apple. If making a decision based on such an interpersonal comparison of preferences, Alice is then said to empathize with Bob.

It is worth noting that, when empathizing with Bob, Alice must always separate her preferences as being herself from her preferences as being Bob. According to Binmore in Binmore [1994, 2005], if Alice identifies so strongly with Bob that she forgets her own preferences (as being herself), then she is instead said to sympathize with Bob.
Furthermore, Binmore justifies the natural development of such empathetic preferences by the obvious need for some equilibrium selection mechanism in many interactive situations that regularly occur in human societies. In order to formally illustrate to what extent such preferences could be used to reach cooperation (e.g., in the game in Figure 6.1 from Section 6.1), let us first define a game with empathetic preferences by extending the notion of strategic game as defined in Definition 2.1 from Chapter 2 (see Section 2.2.1).

**Definition 6.11 (Game with Empathetic Preferences)** A game with empathetic preferences is a tuple $EM = \langle Agt, \{S_i| i \in Agt\}, \{U_i| i \in Agt\}, \{U_{i,j}^E|i,j \in Agt\} \rangle$ where:

- $\langle Agt, \{S_i| i \in Agt\}, \{U_i| i \in Agt\} \rangle$ is a strategic game as defined in Definition 2.1 from Chapter 2 (see Section 2.2.1);
- $U_{i,j}^E : S \to \mathbb{R}$ is a total function defining agent $i$’s empathetic utility for being agent $j$ such that:

  $C4$ there exists $\alpha \in \mathbb{R}_+^*$ and $\beta \in \mathbb{R}$ such that, for every $s \in S$, $U_{i,j}^E(s) = \alpha \times U_j(s) + \beta$.

For any strategy profile $s \in S$, which may be interpreted as a social contract in this game, and for any pair of players $i, j \in Agt$, $U_{i,j}^E(s)$ represents the utility of $s$ for agent $i$ when playing as if he was agent $j$ (when $i$ imagines himself to be agent $j$). One should note that, because of Constraint $C4$, $U_{i,j}^E(s)$ is simply a linear transformation of $U_j$. Therefore for every $s, s' \in S$, we have $U_{i,j}^E(s) \leq U_{i,j}^E(s')$ if and only if $U_j(s) \leq U_j(s')$. The special case where $U_{i,j}^E = U_j$ for every $j \in J$ simply expresses agent $i$’s indifference for being either player within the group $J \in 2^{Agt^*}$ (in other words, $i$ empathizes equally with every member of $J$).

Binmore argues that an empathetic agent $i$ (as if $i$ was in the original position) imagines himself to be in any other agent’s position and seeks to maximize the utility function determined by a combination of his empathetic preferences. According to Binmore, there are two possible ways of defining such a utility function: the approach based on Harsanyi’s view of the aggregation of individual utilities ([Binmore, 1994, p. 293]) or the approach based on Rawls’s view ([Binmore, 1994, p. 295]).

If we follow Harsanyi (Harsanyi [1986]), then we should assume that the agent $i$ assigns equal probabilities to the fact of playing as any other agent, as shown through function $U_i^H$:

$$U_i^H(s) = \frac{1}{|Agt|} \cdot \sum_{j \in Agt} U_{i,j}^E(s)$$  \hfill (6.2)
If we follow Rawls (Rawls [1971]), then we should assume that agent $i$ is not able to attach probabilities to the fact of playing as a given agent in $Agt$. Instead of maximizing the total empathetic utilities of all agents in $Agt$, agent $i$ will seek to maximize the empathetic utility for being the worst-off individual in $Agt$, as shown through function $U^R_i$:

$$U^R_i(s) = \min_{j \in Agt} U^E_{i,j}(s)$$  (6.3)

It is worth noting that both the principle of classical utilitarianism and the maximin criterion, as introduced in the previous sections, can be obtained by considering respectively the functions $U^H_i$ and $U^R_i$ where $U^E_{i,j} = U_j$ for every $j \in Agt$.

Moreover, while determining which utility function one should follow in the original position clearly remains an open question, one can fairly state that the answer simply depends on the context.

Binmore further argues in Binmore [1994, 1998, 2005] that individuals naturally tend to share the same empathetic preferences behind the veil of ignorance: “Insofar as people from similar cultural backgrounds have similar empathetic preferences, it is because the use of the original position in this way creates evolutionary pressures that tend to favor some empathetic preferences at the expense of others” ([Binmore, 1998, p. 178]). As a consequence, it appears that any strategic issue becomes irrelevant in the original position: while assuming that others share the same empathetic preferences in the original position, every player $i$ will simply choose whichever solution $s \in S$ maximizes the empathetic utility function $U^H_i(s)$ or $U^R_i(s)$ (Binmore [1994]). Following this interpretation, we specify a player’s empathetic behavior as in Definitions 6.12 and 6.13.

**Definition 6.12 (Empathetic behavior à la Harsanyi)** In a game with empathetic preferences $EM = \langle Agt, \{S_i| i \in Agt\}, \{U_i| i \in Agt\}, \{U^E_{i,j}| i, j \in Agt\} \rangle$, if an agent $i \in Agt$ empathizes with every other player from $Agt$, then $i$ chooses to perform strategy $s_i$ such that:

$$s \in \arg\max_{s' \in S} U^H_i(s')$$

where $U^H_i$ is defined according to Equation (6.2).

**Definition 6.13 (Empathetic behavior à la Rawls)** In a game with empathetic preferences $EM = \langle Agt, \{S_i| i \in Agt\}, \{U_i| i \in Agt\}, \{U^E_{i,j}| i, j \in Agt\} \rangle$, if an agent

---

1In Binmore [1998, 2005], Binmore proposes a theory, which states that Harsanyi’s utilitarian function only makes sense under the control of some real external enforcement agency (e.g., an all-powerful government), whereas Rawls’ egalitarian function is ideal in the absence of any such external enforcement (Rawls’ egalitarian solution becomes self-enforcing in this case).
$i \in \text{Agt}$ empathizes with every other player from $\text{Agt}$, then $i$ chooses to perform strategy $s_i$ such that:

$$s \in \arg \max_{s' \in S} U^R_i(s')$$

where $U^R_i$ is defined according to Equation (6.3).

When applied to the Hi-Lo game and the Stag-Hunt game from Section 6.1, one can observe that every game with aligned empathetic preferences based on Figure 6.1 (where $0 \leq z < x$) has a unique way to empathize (à la Harsanyi or à la Rawls), which leads the players to reach the $(C,C)$ outcome. More specifically, the corresponding game with aligned empathetic preferences (where $a$ and $b$ respectively stand for Alice and Bob) can be defined as follows:

$$EM = \langle \{a, b\}, \{S_a, S_b\}, \{U_a, U_b\}, \{U^E_{i,j} | i, j \in \{a, b\}\} \rangle$$

with $S_a = S_b = \{C, D\}, U_a(C, C) = U_b(C, C) = x, U_a(D, D) = U_b(D, D) = y, U_a(D, C) = U_b(C, D) = z, U_a(C, D) = U_b(D, C) = 0,$ and $0 < y < x$. Thus, we have that, whenever $0 \leq z < x$, for any setting of the functions $U^E_{i,j}$ in $EM$, any empathetic player (according to either Definition 6.12 or Definition 6.13) will select option $C$. On the other hand, in the case of the Prisoner’s Dilemma, (where $z \geq x$), one should note that some agent $i$’s empathetic behavior depends on the setting of $i$’s empathetic utility functions $U^E_{i,j}$: for example, if $U^E_{i,a}(s) >> U^E_{i,b}(s)$ for every $s \in S$ (i.e., $i$ always largely prefers to be $a$ than $b$), $i$’s empathetic behavior à la Harsanyi will be to choose $D$. Note that, while such an interpretation is clearly counter-intuitive (one would indeed intuitively expect empathetic individuals to only select $C$ in the prisoner’s dilemma), Binmore provides a more precise theory of determining “reasonable” empathetic preferences (i.e., functions $U^E_{i,j}$) in Binmore [1994, 1998].

However, although Binmore’s theory is highly relevant to explain cooperation in various types of social situations, it also has some limitations regarding its connection with strategic thinking. More precisely, his theory does clearly not allow to explain under which conditions individuals will consider their individual preferences and play strategically, and under which conditions they will consider their empathetic preferences and play according to either Definition 6.12 or Definition 6.13. In other words, such a model does not provide an interpretation of the strength with which a particular individual (more or less) empathizes with another person. This remark is even more striking when considering larger games with more than two individuals. In fact, in this type of interactions, an agent may reasonably strictly empathize with some individual(s) while reasoning strategically with the other(s) (see Section 6.3.4 for a concrete illustration). In fact, Binmore’s theory of empathetic preferences does not allow to model this kind of complex situations because it does not incorporate strategic reasoning between competing coalitions.
However, it is worth noting that Binmore’s concept of empathy is closely related to our theory of social ties introduced in Section 6.3. In fact, it appears that an empathetic behavior as in Definitions 6.12 and 6.13 corresponds to solving a decision problem in the context of a strong social tie with the group of all agents (see Definition 6.6 from Section 6.3). However, as our model of ties incorporates strategic reasoning, it clearly allows to resolve the above issues related to Binmore’s model.

6.5 Conclusion

In this work, we have provided an analysis of a well known economic theory that allows to explain collective behavior in situations that involve co-operation: Bacharach’s theory of team reasoning. After discussing the limitations of this theory, we have presented our theory of social ties, which, by generalizing the model introduced in Chapter 4, appears to be well suited to formalize collective behavior in the context of complex strategic interactions. The advantages of our model compared to Bacharach’s theory have been highlighted: while Bacharach’s theory solely relies on binary group identification, social ties require the more general concept of gradual group identification, as already suggested in Chapter 4. More precisely, the proposed comparative analysis has shown that considering this assumption allows us to formalize more intuitively complex types of social interactions that can involve the formation of various coalitions. In fact, suppose that two of my best friends are involved in a fight against each other (which obviously implies the existence of a weak social tie between them). Clearly, in this particular scenario, identifying with the whole group is out of the question since I know that my friends are not willing to collaborate with each other. Our theory of social ties then suggests that I should identify simultaneously with two subgroups (each of which includes only one of my friends) so that I can rationally choose whichever action allows me to resolve the conflict between them, as an attempt to avoid any further damage on the welfare of these subgroups (and consequently on each friend’s own welfare). On the other hand, Bacharach’s theory suggests that I should instead team up with one of my friends and fight against the other (as a result of identifying with either one subgroup or the other), which clearly appears to be overly unrealistic.

Furthermore, we have shown through this study that our theory of social ties could also interpret some empathetic behavior as it shares some common intuitions with Binmore’s own theory of empathetic preferences.
Chapter 7

Summary and Future Research

“Selfishness is not living as one wishes to live, it is asking others to live as one wishes to live.”

— Oscar Wilde

The Soul of Man Under Socialism (1891)

“I believe in one thing - that only a life lived for others is a life worth living.”

— Albert Einstein

(1948)

For many years, it has been assumed that rational behavior was solely determined by one’s individual preferences and one’s ability to think logically. In the line of more recent work, we have argued, through this dissertation, that other relevant factors are also necessary. Among those, it is clear that knowledge and beliefs play an important role in determining one’s rational choice in social interactions. While there indeed exist many different sources of uncertainty (e.g., about each individual’s preferences, past and future moves, available strategies), our main focus in Chapter 3, was to strictly investigate what individuals know and ignore about each other’s future behavior, and how their prediction about what will happen does affect their own behavior (under the assumption that everything else about the game structure, including the players’ preferences, is commonly known). The main result of such a study is that it provides an intuitive explanation for why people deviate from the most optimal behavior. While it is sometimes argued that such behavior is the result of irrationality, we here claim that it is often compatible with some bounded rationality principle. In fact, most human beings are unable to accurately reason about complex interactive knowledge and beliefs (e.g., about what I know that you know that I know that you know that ...). In the same
way, people have also bounded memories, which prevent them to perfectly recall what previously happened and use that information in predicting what will happen next. However, while this kind of cognitive limitations clearly allows to justify such non optimal behavior, its most interesting consequence lies in the fact that it is commonly believed by all individuals involved. Indeed, even though I may be able to perfectly reason about very complex statements, if I do not believe that others share the same ability, then all that cognitive power simply becomes useless as it may then be in my best interest to deviate from the most optimal solution (e.g., the centipede game discussed in Chapter 3 shows that assuming common knowledge of rationality can sometimes lead to some poor outcome).

As another principal component that drives rational decisions in social interactions, we claim through this dissertation that social relationships existing between individuals play a major role in promoting cooperation. More precisely, we first hypothesize, in Chapter 4, that such social ties are used to shape our preferences in a way that simplifies decision making. We show that, through this theory, strong social ties can remove any strategic component from the original situation by simply transforming the individuals’ preferences. As a means to compare this interpretation with other relevant economic theories of social preferences, we introduced a particular two-player coordination game that allows to disentangle between their predictions. We also investigated the well known alternative theory of team reasoning, which relies on group identification, and showed that it is not adequate to model social ties as continuous variables. More generally, this analysis shows that cooperation is compatible with a rational behavior in the context of social ties.

We then validated this theory through an experimental study in Chapter 5, which involved participants who share some existing social ties with one another. While this study shows that increasing the strength of social ties improves the level of coordination by playing fair, it also reveals that such cooperative behavior relies on a combination of a subjective measure of ties (i.e., what they believe to be the social ties) as well as an objective measure of ties (i.e., what the social ties actually are). More precisely, our study shows that being overly optimistic in terms of the expected value of a social tie (i.e., overestimating the strength of a social tie) does not only benefit the welfare of the group but can also serve each individual’s own best interest in situations involving coordination. In fact, it appears that being self-regarding in coordination games similar to those studied in Chapter 5 is the worst possible strategy one could follow to maximize one’s own payoffs.

Finally, through the last chapter of this dissertation (Chapter 8), we have explored this theory of social ties in more details by considering its interpretation in complex social situations that can often be found in various human societies. In fact, societies are made of large amounts of interconnections between individuals,
which result in the formation of groups that may compete with one another. We have therefore shown that our model of social ties provides an intuitive and alternative way to represent such situations where a given individual does, at the same time, (1) cooperate with other members of the group(s) he identifies with, and (2) compete with members of other groups he does not identify with. Through a formal comparison with Bacharach’s team reasoning, we also illustrated the relevance of our model by considering decision making in the particular case of multiple gradual group identifications (which are not allowed in Bacharach’s theory). More generally, our study suggests that, when making a social decision, people naturally evaluate the quality of the links existing between all individuals that they come to interact with. In other words, one’s perception of the structure of a society (through its interconnections) has a direct impact on one’s contribution to the promotion of that society (by maximizing various group utilities).

However, while we believe that the study presented through this dissertation allows to clarify the concept of social rationality, we also realize that it only represents a premise to a multitude of related research paths that should deserve particular attention in the future. Let us provide a few examples here.

Although Chapter 3 has illustrated the relevance of using modal logic as an alternative powerful tool to formally model strategic reasoning in multi-player interactions, it also suggests some further logical analyses. In fact, while the study proposed here strictly focuses on games with perfect information, extending this analysis to games with imperfect information could similarly offer interesting results. However, the corresponding logical framework would then require introducing additional modalities allowing to reason about past events (which is not allowed in ELEG from Chapter 3 that only reasons about future events). Indeed, in order to realistically investigate the epistemic foundations of relevant solution concepts such as the well known forward induction principle introduced in Chapter 4, it is required to formalize how agents can think about counterfactual situations, that is, about what they could have done and what they would have known. While the concept of forward induction has already been studied in economics (see, e.g., Battigalli and Siniscalchi [2002]; Perea [2010]), those analyses remain purely semantic. Providing a formal language that is sufficiently expressive to explicitly reason about counterfactuals on the syntactic level may therefore be useful to study more precisely the epistemic foundations of rationality.

Moreover, the study of games with incomplete information represents a similarly promising area of research. This type of games represent situations where the agents may be uncertain about each other’s preferences. While such games have already been widely studied by economists, there exists no work that considers any of the recent theories of social preferences from the literature. As an example, considering the theory of social ties introduced in this dissertation, we have
argued that their strength directly determine the individuals’ preferences. In this case, as a means to uncover an individual’s true preferences, one may simply need to adjust one’s beliefs about that individual’s social ties with others. However, it appears that acquiring such accurate beliefs does not appear to be an easy task. In fact, while we assumed in Chapter 4 that the existence of a social tie is commonly believed by the two individuals involved, Chapter 5 suggests that estimating social ties between other individuals is more complex and therefore approximated (see the difference between subjective and objective ties in Chapter 5). In this case, relevant research questions may then include the following: how do my beliefs about ties between others affect my own ties with them? How do my ties with others influence my beliefs about the ties between them? While answering those questions clearly requires some further experimental study, the use of logic may be particularly adequate to formalize its results.

As another perspective of future research, it is clear that a logical analysis of normative reasoning also deserves particular attention. In fact, the experimental study from Chapter 5 reveals that, when people do not think strategically, then they tend to follow what they believe to be the norm, that is, what everybody should do in a given situation (see Section 5.4.2 for more details). In this case, it appears that modal deontic logic is a very adequate tool to formalize such reasoning about obligations (what agents ought to do) and permissions (what agents are allowed to do). As an example, we have proposed, through a separated work (Herzig et al. [2011]), a logical framework that allows to represent and reason about agent interactions in normative systems. More specifically, we have shown that a dynamic logic of propositional assignments was sufficient to express interesting properties related to the dynamics of abilities and permissions (e.g., an agent may allow/prevent/authorize/forbid a person to perform a particular action). However, in order to be able to fully characterize normative behavior as it is observed in Section 5.4.2, it is clear that such a logical framework is not sufficient and should then be combined with a game theoretic setting similar to that introduced in Chapter 3 (or in Lorini [2011]; Lorini and Schwarzentruber [2010]). Such a study would therefore allow to bring some interesting insight to the existing relation between strategic reasoning and such normative thinking.

Besides the need for such relevant logical analyses, it is also worth noting that the study of social ties presented in this dissertation also suggests another required improvement of the model presented in Chapter 8, which only considers ties with groups (through the level with which each individual identifies with every given group). One may then wonder how such group identification is determined in the first place. Indeed, it is reasonable to assume that a group is made of agents who share some individual ties between each of its members. Expressing the above functions $k_i$ (for every agent $i$) strictly in terms of individual ties however does not
appear to be an easy task: as already shown in Chapter 5 (see Section 5.3.2.1), one may indeed be very close individually to every member of a group without being socially close to the group itself.

In order to illustrate the existing relation between these two different concepts of social ties, let us consider the case of a three player social interaction, in which the corresponding network of social ties can be represented as in Figure 7.1.

Figure 7.1: Individual social ties in a three player game

In Figure 7.1, for every pair of individuals \( i, j \in \{1, 2, 3\} \), the strength of the individual tie between \( i \) and \( j \) can be characterized by \( st_{ij} \in [0, 1] \) (i.e., \( st_{ij} = 0 \) if the tie is extremely weak, and \( st_{ij} = 1 \) if the tie is extremely strong). In this case, it is easy to see that if all three players are extremely close to one another (i.e., \( st_{12} = st_{13} = st_{23} = 1 \)), then it intuitively implies that all individuals will identify with the same group (i.e., for every \( i \in \{1, 2, 3\} \), \( k_i(\{1, 2, 3\}) = 1 \)). Similarly, if they are not tied with one another (i.e., \( st_{12} = st_{13} = st_{23} = 0 \)), then it intuitively implies that all players will be self-regarding (i.e., for every \( i \in \{1, 2, 3\} \), \( k_i(\{i\}) = 1 \)). However, expressing group identification may become less straightforward in some alternative situation where player 1 is extremely close to both players 2 and 3 (i.e., \( st_{12} = st_{13} = 1 \)) while players 2 and 3 are extremely weakly tied with each other (i.e., \( st_{23} = 0 \)). In this case, a particularly intuitive interpretation would be that player 1 extremely identifies with the two subgroups (i.e., \( k_1(\{1, 2\}) = k_1(\{1, 3\}) = 0.5 \), while both players 2 and 3 are partly self-regarding (i.e., \( k_2(\{2\}) = k_3(\{3\}) = 0.5 \)). In this case, one should note that the example presented in Section 6.3.4 from Chapter 8 allows to illustrate this very plausible interpretation.

More generally, the aim of this refinement of the social ties model from Chapter 8 is to provide an economic analysis of the impact of group identification on strategic interactions within social networks. It is worth noting that this work should also include some comparative analyses with various existing economic models from the recently growing literature on social networks (see, e.g., Bramoullé et al.).
Other related directions of research concern the analysis of social ties and collective reasoning in the context of sequential interactions. As such sequential games can be represented in strategic form (as shown in Section 2.2.2 from Chapter 2), one can therefore apply the model of social ties from Chapter 8 to specify both individualistic and cooperative behavior. For example, it allows to explain mutual cooperation in the well known trust game (see Figure 2.8 in Chapter 2) as well as in the ultimatum game (see Figure 2.2 in Chapter 2). However, one should note that this theory cannot capture the type of apparently irrational behavior elicited in the latter game. Indeed, experimental studies have shown that people sometimes appear to select an option that is bad for themselves as well as for the group, in response to a prior behavior of their partner that they judge to be unfair (see, e.g., Falk et al. [2003]). Our theory of social ties is clearly unable to express such negative reciprocity or punishment. Yet, such behavior appears to be of high relevance to the concept of social ties. One may intuitively expect that more severe punishments are performed when strong social ties are involved (e.g., the higher the tie between two individuals, the more important the impact of betrayal of one towards the other). Some further experimental study would therefore be required to investigate this issue.

Moreover, such an interpretation, combined with the quantitative aspect of social ties (i.e., the fact that two individuals can be more or less tied with each other), clearly suggests a dynamic characterization of those ties, which we did not consider here. In fact, one may reasonably assume that cooperative behavior allows to strengthen a tie between individuals while individualistic behavior simply weakens it. While there exists no clear evidence in the literature that supports this intuitive claim, it therefore requires some further experimental study that investigates social ties in the context of dynamic interactive situations (i.e., in sequential games or repeated games).

Similarly, one should note that the comparative theoretical analysis of our model of social ties with Bacharach’s theory of team reasoning from Chapter 8 also suggests some further relevant experimental work. Indeed, while the coordination game introduced in Chapter 4 does not allow to disentangle predictions made by the two theories, the games presented in Section 6.3.3 from Chapter 8 illustrate some clear disagreement between both models. Testing the validity of both theories would therefore require testing experimentally those games, which would also improve our general understanding of social ties.

Furthermore, it is worth noting that the work presented in this dissertation remains ambiguously regarding what characterizes a fair solution for every member of a group. Through Chapters 4 and 8, we consider two different approaches to compute collective payoff: either one follows Rawls’ criterion of fairness and aims
at maximizing the worst-off individual within the group, or one follows classical utilitarianism and aims at maximizing the global efficiency of the group. However, one might wonder which of these methods an individual should actually follow. While it is fair to state that this choice depends on the context, Binmore proposes a theory in Binmore [1998, 2005] that relies on Rawls’ concept of the original position (see Section 6.4 from Chapter 8). Such a theory states that classical utilitarianism only makes sense under the control of some real external enforcement agency (e.g., an all-powerful government). Binmore indeed argues that such an enforcement is necessary to ensure that any fair decision that is taken in the original position, based on global efficiency, is actually followed when the veil of ignorance is removed and roles are randomly assigned to individuals. In other words, without any such enforcement, individuals that are assigned roles of disadvantaged players may then deviate from the utilitarian solution they committed to in the original position (i.e., what looks fair in the original position does not necessarily look fair when one gets to know his own true identity), which may eventually result in damaging the collective welfare. On the other hand, Binmore also states that Rawls’ criterion of fairness is ideal in the absence of any external enforcement. In this case, as an egalitarian decision is taken in the original position, removing the veil of ignorance cannot create any incentives for deviating from it. In fact, such a decision already maximizes the interests of those individuals that are assigned roles of disadvantaged players (i.e., what looks fair in the original position also necessarily looks fair when one gets to know his own identity). In that sense, Rawls’ egalitarian solution is simply self-enforcing.

As a means to verify such a reasonable theory and investigate other plausible hypotheses that could clarify which collective payoff function should apply to the above model of social ties and in which context, the use of computer agent-based simulations appears as an obvious choice. However, while in most such social simulations, interactions between agents rely on some random matching process (i.e., an agent has the same probability to interact with any other agent), one could reasonably assume that people have a natural tendency to interact more often with individuals they are strongly tied with. This assumption is also supported by some recent work in Alger and Weibull [2012], which suggests that one’s degree of morality somehow corresponds to the probability of interacting with individuals who share the same preferences. Therefore, in addition to bringing some more insight to the determination of the collective payoff function within our model of social ties, such simulations may also become particularly useful to study the dynamic properties of social relationships (e.g., when does it pay off to increase/decrease one’s social tie with an individual?).
Chapter 8

Résumé en français

8.1 Introduction

Que signifie être rationnel? Il est juste de dire que cette simple question a soulevé de nombreux débats au cours des dernières années à travers de nombreuses disciplines parmi lesquelles la philosophie, l’économie, la psychologie, et la sociologie. Néanmoins, malgré la difficulté pour les scientifiques de fournir une définition claire, le terme “rationnel” est excessivement utilisé dans les conversations quotidiennes, ce qui indique que les gens ont au moins une idée générale de son concept. En effet, chacun conviendra que, d’une manière générale, la rationalité évoque le fait de penser et de se comporter rationnellement ou logiquement. Cependant, une si simple définition reste largement ambiguë puisqu’elle repousse simplement le problème: que signifie penser et se comporter rationnellement? Afin de donner une réponse plus précise à cette question et d’étudier le rôle qu’un tel principe joue dans la coopération humaine, cette thèse combine des approches et des méthodologies différentes provenant de l’informatique (la logique) et l’économie (la théorie des jeux, l’expérimentation). Cependant, dans le but de justifier la besoin pour une telle étude interdisciplinaire, il est utile de distinguer d’abord les concepts de raison et de rationalité: alors que la raison peut être définie à travers la capacité psychologique à établir et à vérifier des faits à partir d’informations perçus, la rationalité concerne plutôt la procédure d’optimisation de choix. Dans ce cas, puisque de tel choix conscients résultent clairement d’une réflexion intérieure, one peut affirmer que la rationalité implique naturellement l’usage de la raison.

De plus, il est coutume parmi les philosophes de distinguer les notions de rationalité théorique et rationalité pratique (voir, e.g., Kalberg [1980]). D’une part, la rationalité théorique, qui repose sur une argumentation et des preuves, traite simplement de la régulation de nos propres croyances. En revanche, la

\[1\] De manière plus rigoureuse, l’acceptance devrait être distinguée de la croyances, comme
rationalité pratique correspond à la stratégie de vivre la meilleure vie possible, atteindre ses objectifs les plus importants, et maximiser ses propres préférences autant que possible.

Afin d’illustrer ces concepts philosophiques, supposons que je sais que fumer abondamment tue, et que cela implique que fumer va, de manière certaine, détériorer ma santé physique. En assumant que je préfère rester en vie et en bonne santé aussi longtemps que possible, je serais alors pratiquement irrationnel de fumer ne serait-ce qu’une seule cigarette car ce simple choix ne me permettrait pas d’être dans le meilleur état de santé possible (i.e., ce choix n’est pas optimal). Par contre, fumer en étant pratiquement rationnel exigerait la reconsidération soit des mes croyances (e.g., finalement, je ne crois pas que fumer va détériorer ma santé), soit mes préférences (e.g., finalement, il n’est pas important pour moi de vivre une longue vie). En revanche, il peut être théoriquement rationnel pour moi de croire que fumer une seule cigarette n’aura qu’un effet négligeable sur ma santé. Cependant, il faut noter qu’il serait théoriquement irrationnel de croire que cela ne va pas du tout détériorer ma santé (ceci serait incompatible avec ma croyance initiale de la déclaration contraire).

Ce simple exemple intuitif clarifie le fait que les êtres humains ne sont pas rationnels par définition, mais ils peuvent penser et se comporter rationnellement ou pas, selon qu’ils appliquent les stratégies de la rationalité théorique et de la rationalité pratique aux pensées qu’ils acceptent ainsi qu’aux actions qu’ils accomplissent. Par ailleurs, ils est à noter que ces concepts se soutiennent mutuellement. De fait, alors que la rationalité théorique m’aide à accomplir mes buts pratiques, la rationalité pratique peux me permettre d’améliorer la qualité de mes croyances. Compte tenu de la forte pertinence de chacune de ces notions de rationalité par rapport au fonctionnement du comportement humain, ils caractérisent alors le thème principal dont fait l’objet cette thèse.

Au cours des dernières décades, étudier le rôle que la rationalité pratique joue réellement dans la monde social est devenu l’objectif principal de nombreux économistes qui ont utilisé les outils mathématiques pour modéliser la prise de décision. Cette intérêt a naturellement mené au développement de la théorie des jeux, qui représente l’étude de la prise de décision stratégique à travers des modèles mathématiques de conflits et de coopération entre agents intelligents rationnels (Gintis [2000]; Myerson [1997]; Osborne and Rubinstein [1994]; Osborne [2004]). Le concept d’un jeu dont cette théorie fait référence peut théoriquement représenter n’importe quelle sorte d’interaction sociale. Formellement, un jeu (non-coopératif) 1 consiste en un ensemble de joueurs, un ensemble d’actions (ou argumenté dans Tuomela [2000]; bien que chacun de ces concepts représente des états cognitifs, les croyances sont involontaires (i.e., elles ne sont pas sujets à un contrôle direct volontaire), alors que les acceptances sont volontaire et intentionnelles.

1Un jeu est non coopératif dans le sens où il représente un modèle détaillé de toutes les actions
stratégies) disponibles pour ces joueurs, et une spécification des récompenses (ou conséquences) pour chaque combinaison d’actions. Il convient de noter que, bien que la théorie des jeux a été principalement ancrée en économie, elle est désormais largement utilisée et étudiée dans d’autres domaines tels que les sciences politiques, la psychologie, ainsi que la philosophie et la biologie. Notamment, l’intérêt pour cette théorie s’est étendu plus récemment au domaine de l’informatique à travers le développement grandissant de l’intelligence artificielle et les systèmes multi-agents. Les applications concrètes sont nombreuses et incluent la robotique, les réseaux de communication électronique (e.g., le commerce électronique), l’éducation interactive, le divertissement interactif (i.e., les interactions humains-ordinateurs), et la résolution des problèmes de sûreté et de sécurité. Dans chacune de ces situations, afin pour ces agents artificiels d’interagir efficacement avec des êtres humains, ils doivent clairement être capable de comprendre les principes de base du comportement social, tels qu’ils apparaissent dans les sociétés humaines. Une telle problématique suggère alors fortement le besoin de définir une théorie formelle de la rationalité dans le contexte d’interactions sociales.

Cependant, malgré son importance indéniable envers l’étude de la rationalité pratique, la théorie des jeux classique n’apparaît pas comme l’outil le plus efficace pour analyser les connections fondamentales avec la rationalité théorique. En effet, comme suggéré précédemment, la rationalité théorique consiste simplement à suivre une manière de raisonner consistante et optimale, et la théorie des jeux ne fournit pas un langage suffisamment riche qui permette de modéliser de façon non ambiguë ce type de raisonnement. Afin de répondre aux besoins de formaliser une réflexion logique, l’utilisation de la logique propositionnelle est souvent privilégiée par les informaticiens. Un tel système formel concerne en effet le raisonnement sur des propositions, chacune représentant simplement une description possible du monde: par exemple, étant donné deux propositions \( p \) et \( q \), si \( p \) est vrai et il est le cas que “si \( p \) est vrai, alors \( q \) est vrai”, alors on peut en déduire que \( q \) est également vrai (cette règle d’inférence est connu sous le nom de Modus Ponens). Néanmoins, en contrepartie d’être extrêmement simple, une telle logique est insuffisante pour exprimer les différents états mentaux tels qu’ils peuvent apparaître dans les interactions sociales. À cet effet, un autre système formel a été introduit pour étendre cette logique avec l’addition d’opérateurs supplémentaires exprimant des modalités du type “il est possible/nécessaire que …”, “il est permis/obligatoire que …”, “je crois/sais que …”, etc... Ces opérateurs modaux supplémentaires sont en effet particulièrement intéressant à l’étude de la rationalité car ils permettent de formaliser le raisonnement sur les attitudes mentales des agents. Par exemple, il
est possible de définir la règle suivante dans un tel langage: si l’individu \( i \) croit que la proposition \( p \) est vrai, alors \( i \) croit que \( i \) croit que \( p \) est vrai (cette règle est généralement connu sous le nom de introspection positive). Ce type de logique, qui est connu sous le nom de logique modale (Blackburn et al. [2002]; Chellas [1980]; Hintikka [1962]; Hughes and Cresswell [1968]), est également souvent associée à d’autres logiques telles que la logique épistémique (raisonnement sur la connaissance), la logique temporelle (raisonnement sur le temps), la logique déontique (raisonnement sur les obligations), et la logique dynamique (raisonnement sur les programmes complexes). Plus généralement, les applications de la logique modale sont particulièrement importantes en philosophie, linguistique, et dans les différents domaines de l’informatique tels que l’intelligence artificielle, les systèmes distribués, les bases de données, la vérification de programmes, et la cryptographie. Un objectif de cette thèse est de montrer que cette logique est également particulièrement pertinente dans le domaine de l’économie.

Néanmoins, bien que la combinaison de la théorie des jeux et la logique représente un outil analytique puissant pour analyser les principes essentiels de la rationalité, leur seule limitation concerne la vue largement idéalisée qu’elle procure. Comme suggéré précédemment, la motivation principale pour formaliser la rationalité est de pouvoir correctement prédire le comportement humain dans les interactions sociales. En effet, être rationnel devient inutile si je crois, à tort, que tous les individus sont rationnels, me conduisant ainsi à réaliser de mauvais choix avec potentiellement des conséquences catastrophiques. Il est aujourd’hui largement admis que la théorie des jeux classique n’a effectivement pas réussi à remplir sa motivation originale de prédire correctement la prise de décision humaine, ce qui explique le récent intérêt grandissant pour l’expérimentation économique. Similairement aux science physiques, l’utilisation d’expériences contrôlées s’est trouvée être largement pertinente à l’étude de questions économiques (Camerer [2003]; Roth and Kagel [1995]). Dans le but de tester la validité de certaines théories économiques, de telles expériences utilisent généralement de l’argent pour motiver les sujets humains afin de mimer les incitations existantes de la vie de tous les jours. De telles méthodes empiriques permettent d’explorer des concepts très importants parmi lesquels l’altruisme, la réciprocité, et les émotions, qui ont tous, pour longtemps, été ignorés par la théorie économique classique. De fait, de nombreuses preuves empiriques ont déjà suggéré que les êtres humains considèrent naturellement le bien-être des autres dans les interactions sociales. Cependant, alors que ces études ont permis de clarifier la manière avec laquelle chaque individu peut contribuer à la promotion de la société (e.g., à travers un comportement juste et coopératif), il reste à déterminer l’impact qu’une société peut avoir sur le comportement de ces membres. En effet, toute société humaine est constituée d’un groupe d’individus liés les uns aux autres à travers des relations sociales plus ou moins persistantes.
En conséquence, il est raisonnable d’affirmer que l’environnement d’une personne est largement responsable à la détermination de son propre bien-être. Après tout, la contribution d’un individu au bien-être d’une société a pour principal (si ce n’est le seul) but d’améliorer sa propre qualité de vie au sein de cette société.

La contribution principale de cette thèse est alors d’analyser ce problème par l’étude de certains aspects cruciaux de la rationalité humaine telle qu’elle apparaît dans les interactions sociales. Notre objectif est d’expliquer comment ces facteurs peuvent déterminer un comportement social qui promeut le bien-être de la société entière. Dans le but d’obtenir une définition formelle et réaliste du concept complexe de la rationalité sociale, cette étude repose sur les différentes méthodologies qui ont été précédemment citées: la théorie des jeux, la logique modale, et l’expérimentation économique. Plus spécifiquement, nous soutenons que les composantes suivantes sont essentielles à la définition de la rationalité sociale dans les situations d’interaction sociale:

- **Les préférences individuelles.** Puisque la rationalité pratique traite de l’optimisation de buts, il est nécessaire de définir ce qui caractérise la meilleure solution pour chaque individu. L’approche que nous suivons ici est quantitative à ce sujet, c’est-à-dire, elle repose sur la mesure de l’utilité pour chaque conséquence possible.

- **Les connaissances et croyances.** Il est clair que la rationalité théorique dépend largement de l’état épistémique d’un agent. En considérant l’exemple ci-dessus, si je ne sais pas que fumer tue, alors il peut être rationnel pour moi de fumer. Dans le contexte des interactions sociales, le problème devient même plus complexe puisque tous les agents doivent alors considérer ce qu’ils savent à propos de ce que chacun sait.

- **Les liens sociaux entre individus.** Ce facteur n’est généralement pas mentionné dans la littérature pour définir la rationalité. Nous affirmons ici que le type de relations sociales qui peut exister entre individus impliqués dans certaines interactions sociales a un effet sur leur rationalité respective. Plus précisément, nous défendons le fait que le niveau avec lequel chaque agent est lié avec chaque autre agent affecte directement les préférences individuelles, et par conséquent le comportement.

### 8.2 Rationalités dans les interactions sociales

Ce chapitre fournit une brève introduction concernant les théories formelles existantes qui caractérisent le concept de rationalité. Plus précisément, nous présentons la théories de l’utilité ainsi que la théorie des jeux dans le but de représenter
les préférences rationnelles ainsi que la prise de décision rationnelle dans le contexte d’interactions stratégiques sociales. Afin de raisonner sur la connaissance et les croyances, nous présentons la logique modale épistémique et illustrons la puissance descriptive de cette méthode par un simple exemple. De plus, nous démontrons la nécessité d’étendre la théorie des jeux classique de manière à ce qu’elle puisse incorporer les types de raisonnement suivis par les sujets humains dans de nombreuses expériences économiques. Nous défendons alors le fait que, souvent, les preuves apparentes de comportement irrationnel ne contredisent pas l’hypothèse classique de la rationalité.

8.3 La rationalité épistémique dans les jeux extensifs

A travers ce chapitre, nous étudions les fondations épistémiques de la rationalité grâce à une analyse logique des interactions sociales séquentielles. Bien que des travaux similaires aient déjà été réalisés en économie (e.g., Aumann [1995, 1999a]; Aumann and Brandenburger [1995]; Battigalli and Siniscalchi [1999, 2002]), la particularité de cette analyse est qu’elle repose sur la logique modale. En effet, nous argumentons que la logique modale est un outil inestimable qui permet des analyses détaillées tout en exprimant des concepts qui sont informellement ou vaguement capturés par le langage classique de la théorie des jeux. En particulier, nous montrons qu’un tel outil formel est idéal pour modéliser de manière non ambiguë les connaissances d’agents à propos de ce que chacun sait. Cette étude fait alors référence au domaine récent appelé “épistémologie interactive formelle” (un terme inventé par Aumann [1999b]), qui traite de la logique de la connaissance et de la croyance lorsqu’il existe plus d’un agent. De plus, l’autre principale caractéristique de ce travail est qu’il considère la dimension temporelle de la rationalité qui est souvent ignorée dans la littérature. Nous montrons alors que la grande capacité expressive de la logique proposée permet d’améliorer la compréhension de la relation existante entre le temps et la connaissance dans les interactions sociales. Afin d’illustrer l’utilisation d’une telle logique, nous fournissons une preuve syntaxique du théorème bien connu d’Aumann indiquant que l’induction rétrograde dans les jeux à information parfaite peut être dérivée à partir de l’hypothèse de connaissance commune de rationalité mutuelle (Aumann [1995]). La définition de rationalité qui est considéré dans ce théorème peut être décrite de la manière suivante: “Peut importe où le joueur se trouve - à quelle point du jeu - il ne choisira pas une stratégie en sachant que cela lui apportera moins que ce qu’il aurait pu obtenir avec une stratégie différente” (Aumann [1995]).

En conséquence, nous montrons qu’une telle étude logique ne permet pas seule-
ment de clairement identifier les hypothèses épistémiques requises qui sont seulement implicites dans la preuve originale d’Aumann, mais elle permet également d’affaiblir l’énoncé original du théorème et de répondre à des questions pertinentes liées aux mécanismes d’apprentissage et de mémorisation, aux introspections positives et négatives, au raisonnement temporel et à la rationalité limitée. De plus, nous montrons qu’une telle analyse permet également d’obtenir une réponse formelle à la critique principale du théorème d’Aumann dans la littérature de la théorie des jeux (voir Stalnaker [1998]).

8.4 Les liens sociaux et la coordination stratégique

Dans les théories économiques classiques, la plupart des modèles assume que les agents sont individualistes et maximisent leur propre profit matériel. Cependant, comme mentionné précédemment, d’importantes preuves expérimentales en économie et psychologie sociale ont démontré l’existence de déviations persistantes d’un tel comportement individualiste dans de nombreuses situations stratégiques. Ces résultats suggèrent le besoin d’incorporer les préférences sociales dans les modèles de la théorie des jeux. De telles préférences décrivent le fait qu’un joueur ne considère pas seulement son propre profit matériel, mais également celui des autres joueurs (Margolis [1982]). Les différentes normes sociales créées par l’environnement culturel dans lequel les êtres humains vivent apportent des précisions concernant la manière dont de telles données expérimentales peuvent être interprétées: la justice, l’aversion à l’inégalité, la réciprocité, et la maximisation du bien-être social sont des concepts bien connus des économistes comportementaux qui jouent un rôle important dans la prise de décision interactive (e.g., voir Charness and Rabin [2002]; Fehr and Schmidt [1999]; Rabin [1993a]).

En effet, de nombreux jeux économiques simples, tels que le jeu de confiance (Berg et al. [1995]) et le jeu de l’ultimatum (Güth et al. [1982]), ont été largement étudiés dans les années passées parce qu’ils illustrent bien la faiblesse de la théorie des jeux traditionnelle et son hypothèse de rationalité individualiste. De plus, étant donné la simplicité de tels jeux, l’argument d’une rationalité limitée (Gigerenzer and Selten [2001]) ne semble pas être suffisant pour justifier le comportement observé. Les préférences sociales apparaissent alors comme une option plus réaliste car elles permettent d’expliquer les comportement résultant tout en considérant des agents rationnels.

Cependant, bien que de nombreuses études expérimentales (e.g., Berg et al. [1995]; Güth et al. [1982]) ont montré que les gens manifestent naturellement des préférences incorporant le profit des autres lors d’interactions avec des parfaits étrangers, on peut s’interroger dans quelle mesure l’existence de relations sociales peut influencer le comportement. L’aspect dynamique des préférences sociales...
semble très lié à celui des liens sociaux: intuitivement, un individu va coopérer plus avec un ami qu’avec un étranger, ce qui peut ainsi avoir comme conséquence de renforcer le niveau d’amitié.

L’énoncé suivant caractérise la condition minimale que nous considérons pour l’existence d’un tel lien social entre deux individus:

**Statement 8.4.0.1** Un lien social existe entre deux individus si et seulement si ils partagent les mêmes propriétés sociales qui définissent leur identité sociale, et cela est croyance commune entre eux.

De plus, une propriété important des liens sociaux repose sur son aspect quantitatif, c’est à dire, deux individus peuvent être plus ou moins liés l’un avec l’autre. Plus précisément, nous assumons qu’un lien social entre deux individus peut être mesuré sur une échelle variant de 0 à 1, où 0 et 1 correspondent respectivement à la valeur minimale et la valeur maximale du lien. Selon notre interprétation, la dimension quantitative d’un lien social entre deux individus dépend des variables suivantes:

- La quantité et l’importance des propriétés sociales partagées qui définissent les identités sociales des deux individus.
- La quantité et la qualité des interactions passées entre les deux individus.

Formellement, considérons deux joueurs $i$ et $j$. Soit $S_i$ et $S_j$ respectivement l’ensemble des stratégies de $i$ et $j$, et soit $\pi_i(s_i, s_j)$ la fonction déterminant le profit matériel pour le joueur $i$ lorsque $i$ et $j$ jouent respectivement leur stratégie $s_i$ et $s_j$. Pour chaque $s_i \in S_i$ et $s_j \in S_j$, la fonction d’utilité de *Liens Sociaux* pour le joueur $i$ est donnée par:

$$U_i^{ST}(s_i, s_j) = (1 - k_{ij}) \cdot \pi_i(s_i, s_j) + k_{ij} \cdot \max_{s'_j \in S_j} U(s_i, s'_j)$$

où $k_{ij} \in [0, 1]$ défini le lien social entre $i$ et $j$.

La fonction $U(s_i, s_j)$ correspond à la fonction d’utilité de groupe qui peut (par exemple) être caractérisé par l’un des deux principes suivants bien connus.

Définissons premièrement la fonction d’utilité de groupe $U_m(s_i, s_j)$ qui satisfait le critère de *maximin* de Rawls (*Rawls [1971]*).

$$U_m(s_i, s_j) = \min \{ \pi_i(s_i, s_j), \pi_j(s_i, s_j) \}$$

Ce critère correspond à donner un poids infiniment grand aux bénéfices de la personne qui se trouve dans la situation la moins avantageuse.
Une alternative consiste à considérer une fonction du bien-être social \( U_s(s_i, s_j) \) qui satisfait l’utilitarisme classique (i.e., en maximisant la somme totale des profits de tous les joueurs).

\[
U_s(s_i, s_j) = \pi_i(s_i, s_j) + \pi_j(s_i, s_j)
\]

Dans ce cas, l’interprétation de la fonction d’utilité de Liens Sociaux peut être interprétée de la manière suivante. Mettre \( k_{ij} \) à 0 correspond à un lien inexistant entre les individus (e.g., \( j \) est un parfait étranger pour \( i \)) alors que mettre \( k_{ij} \) à 1 signifie que \( i \) se sent socialement très proche de \( j \) (e.g., \( j \) est le meilleur ami de \( i \)). Dans ce dernier cas, il faut noter que, en présence d’un fort lien social avec l’agent \( j \), l’agent \( i \) ne perçoit plus un problème stratégique: en effet, la stratégie \( s_j \) de \( j \) devient inutile au calcul de l’utilité de \( i \). Ainsi, l’agent \( i \) a seulement besoin de résoudre un problème classique de la prise de décision individuelle en sélectionnant l’action à partir du profil de stratégies qui maximise l’utilité du groupe. En conséquence, l’action de \( i \) peut être interprétée comme “faire ce qui est bien pour le groupe, en assumant que l’autre joueur fait également ce qui est bien pour le groupe”.

En d’autres termes, l’hypothèse principale qui résulte de ce modèle est que les relations sociales peuvent influencer le choix des joueurs en modifiant leurs préférences: un agent peut choisir d’être juste conditionnellement à la proximité relative envers son partenaire. Afin d’examiner cette hypothèse, nous proposons une analyse théorique d’un nouveau genre de jeu de coordination à deux joueurs qui permet de séparer les prédictions de théories basées sur l’intérêt personnel, les préférences sociales, et les liens sociaux. Le jeu correspondant, dénommé le jeu d’entrée, est illustré par la Figure 8.1.

A travers ce jeu, nous démontrons également la nécessité d’introduire le modèle précédent pour capturer le concept de liens sociaux comme des variable continues. En effet, en plus de défendre le fait que les liens sociaux reposent fortement sur l’identification de groupe, nous montrons que les théories de raisonnement en équipe (Bacharach [1999]; Colman et al. [2008]; Hakli et al. [2010]; Sugden [2000, 2003]; Tuomela [2010]) sont trop limitées pour remplir ce rôle car elles sont construites à partir d’une interprétation binaire d’identification de groupe (i.e., soit un individu s’identifie à un groupe ou pas).

Afin d’illustrer les différences entre notre modèle de liens sociaux et le concept de raisonnement par équipe, nous considérons un jeu simple et concret de deux joueurs où un joueur \( i \) peut choisir entre trois options: \( A \), \( B \), et \( C \). Dans un tel scénario, le profit de chaque joueur est déterminé uniquement par ces options selon la Table 8.1 (pour simplicité, le joueur \( j \) n’a aucun contrôle sur le résultat). Il faut noter que la fonction d’utilité de groupe peut alors correspondre soit à l’utilitarisme (i.e., la somme des profits individuels), soit au principe du maximin (i.e., le minimum des profits individuels).

190
Appliquer le raisonnement par équipe à cette situation particulière mène à la prédiction suivante: l’agent $i$ va jouer $A$ s’il raisonne en $I$-mode (le joueur $i$ est individualiste), et le joueur $i$ va jouer $B$ s’il raisonne en $we$-mode (le joueur $i$ s’identifie au groupe). En conséquence, selon les théories de raisonnement en équipe (et indépendamment de la fonction d’utilité de groupe qui peut correspondre à l’utilitarisme ou au principe du maximin), le joueur $i$ ne choisira jamais l’option $C$.

D’un autre coté, considérer le même jeu à travers notre modèle de liens sociaux mène à une interprétation différente: dans ce cas, le joueur $i$ sélectionnera $A$ si les individus sont parfait étrangers (e.g., $k_{ij} = k_{ji} = 0$), et le joueur $i$ sélectionnera $B$ si, par contre, les individus sont extrêmement proches l’un de l’autre (e.g., $k_{ij} = k_{ji} = 1$). Cependant, si les joueurs sont ni très proche, ni très étrangers, mais, par exemple, de simples connaissances (e.g., $k_{ij} = k_{ji} = 0.5$), alors le joueur $i$ choisira $C$ (en assumant soit l’utilitarisme, soit le principe du maximin pour la fonction d’utilité de groupe), comme un compromis entre être strictement individualiste et
strictement motivé par l’intérêt du groupe.

### 8.5 Les effets comportementaux des liens sociaux

Mesurer les effets des relations sociales sur le comportement humain n’est pas nouveau dans les domaines de l’économie et de la psychologie sociale. Dans les années passées, de nombreuses études ont effectivement montré que les gens tendent à coopérer plus avec des individus qui appartiennent au même groupe qu’avec des individus qui appartiennent à des groupes différents (voir, e.g., Brewer [1979, 1999]; Chen and Li [2009]; Tajfel and Turner [1979]; Tajfel et al. [1971]). Ces observations ont mené à la distinction entre le concept de *in-group* qui constitue un groupe social auquel un individu s’identifie psychologiquement comme un de ses membres, et celui de *out-group* qui, au contraire, représente un groupe auquel un individu ne s’identifie pas comme un de ses membres. Alors qu’un tel comportement de *in-group* peut raisonnablement être induit par une large variété de phénomènes tels que la culture, la religion, le genre, et la race, le paradigme bien connu du group minimal (Tajfel [1970]) suggère que même des caractéristiques sans aucune signification peuvent suffire à déclencher une identification de groupe. Cependant, malgré les preuves empiriques pour un favoritisme *in-group* (i.e., favoriser des membres de son groupe par rapport aux non-membres), il reste à clarifier ce qui détermine un *in-group* en premier lieu. En fait, on peut facilement imaginer que différents niveaux de *in-group* existent, comme suggéré précédemment: par exemple, un individu peut simultanément s’identifier comme un économiste, un informaticien, et un membre de son club de sport favoris. Notre objectif ici est alors d’examiner l’impact que la variation de ce niveau de *in-group* peut avoir sur le comportement humain: un individu se comporte-t-il de la même manière lorsqu’il interagit avec un ami proche, une simple connaissance, ou un étranger?

Afin de répondre à ce genre de questions, nous proposons une expérience qui concerne deux versions du jeu de coordination asymétrique présenté dans la section précédente (i.e., le jeu d’entrée de la Figure 8.1 avec et sans l’option de sortie). Dans ce contexte, nous varions le niveau de liens sociaux en créant des interactions avec des partenaires de différents *in-groups* (i.e., coéquipiers dans un sport collectif, membres du même club de sport, membres de la même université). L’objectif général de cette étude expérimentale est de vérifier la validité du modèle de liens sociaux présenté dans la section précédente.

De plus, nous utilisons des questionnaires directs afin de mesurer plus précisément les connections sociales entre chaque individu et leur propre équipe sportive. Dans ce cas, nous distinguons deux types différents de liens sociaux: un lien sub-
jectif mesure le niveau avec lequel un individu particulier ressent une connexion avec sa propre équipe (à travers l’évaluation de l’individu à propos des autres membres de l’équipe), alors que un lien objectif mesure le niveau avec lequel une équipe est réellement proche d’un individu (à travers l’évaluation des membres de l’équipe à propos de l’individu). Après avoir comparer les deux types de liens sociaux (subjectif and objectif) l’un avec l’autre, nous examinons alors le rôle que chacun joue dans la détermination du comportement individuel.

Le résultat principal de cette étude est la preuve empirique révélant que des forts liens sociaux aident les gens à se coordonner et à promouvoir le bien-être du groupe. Plus précisément, nos observations supportent le modèle de liens sociaux proposé dans la section précédente, et rejettent les autres théories pertinentes de préférences sociales. De plus, nous montrons que les liens sociaux permettent de promouvoir un sens de la justice dans le contexte des jeux de coordination présentés ci-dessus (i.e., le jeu d’entrée de la Figure 8.1 avec et sans l’option de sortie), ce qui mène les gens à prendre des décisions comme s’ils étaient incapables de distinguer leur propre identité de celle de leur partenaire (cf. la position originale de Rawls). Cette analyse suggère donc que les liens sociaux peuvent être utilisés comme un appareil qui permet d’augmenter le degré d’individualisme (lorsque le lien social diminue), ou d’augmenter le degré de justice (lorsque le lien social augmente), tel que spécifié par notre modèle de liens sociaux.

De plus, cette analyse expérimentale montre que le comportement d’un individu est affecté par son lien social ressenti (subjectif) avec l’autre, ainsi que par son lien social réel (objectif). Plus précisément, nous observons que les gens optomistes qui surestiment la force du lien social tendent à se comporter de manière plus juste en coopérant avec l’autre, alors que les gens pessimistes qui sous-estiment la force du lien social ne s’identifient à aucun groupe et ne parviennent donc pas à se coordonner dans le jeu proposé. Un tel résultat suggère clairement qu’être individualiste ne profite pas toujours dans le contexte des jeux de coordination. En effet, dans ce type d’interaction, il est dans l’intérêt personnel d’un individu d’être lié socialement avec les autres, ce qui permet donc de maximiser ses profits à travers la coordination.

8.6 Vers des sociétés collaboratives

De nombreuses tâches de la vie de tous les jours requièrent les individus d’agir collectivement et de se coordonner afin de poursuivre un but commun. Les exemples incluent les musiciens d’un orchestre qui ont besoin d’agir ensemble d’une manière bien spécifique pour jouer une symphonie particulière, ou les joueurs d’une équipe de football qui doivent se coordonner les uns avec les autres pour pouvoir marquer un but. Même parmi les autres tâches qui sont réalisables par un unique indi-

L’objectif général, ici, est d’utiliser la théorie des jeux pour combler l’écart entre un comportement individuellement égoïste et la coopération sociale dans le contexte d’interactions stratégiques. Dans cette perspective, nous fournissons l’analyse d’une théorie bien connue dans la littérature économique qui explique comment les agents, humains ou artificiels, réussissent à résoudre les problèmes de coordination dans le contexte d’une activité commune: la théorie de raisonnement par équipe de Bacharach. À la suite d’une discussion des limitations de cette théorie dans la modélisation de comportements sociaux intuitifs, nous proposons la généralisation suivante du modèle de liens sociaux présenté précédemment.

Considérons un ensemble de joueurs $Agt$, un ensemble de conséquences $S$, et pour chaque coalition $J \subseteq Agt$ d’au moins un agent, $S_J$ et $U_J$ dénotent respectivement l’ensemble des stratégies pour $J$ et la fonction déterminant le profit matériel pour la coalition $J$ pour chaque conséquence. Pour chaque conséquence $s \in S$, la fonction d’utilité de Liens Sociaux pour le joueur $i$ est donnée par:

$$U_i^{ST}(s) = \sum_{J \subseteq Agt \setminus \{i\}} k_i(J \cup \{i\}) \cdot \max_{s'_{J} \in S_J} U_{J \cup \{i\}}(s_{-J}, s'_{J})$$

où $s_{-J} \in S_{-J}$ dénote la stratégie commune pour la coalition $Agt \setminus J$ (i.e., $S_{-J} = S_{Agt \setminus J}$), et $U_J(s)$ peut être défini selon l’utilitarisme ou le principe de maximin, comme introduit dans la section 8.4.

De plus, la fonction de liens sociaux $k$ est défini de la manière suivante. Pour chaque agent $i \in Agt$ et pour chaque groupe $J = \{G \in 2^{Agts} | i \in G\}$, le paramètre $k_i(J)$ défini la mesure du lien social entre l’agent $i$ et le groupe $J$ dans le contexte du jeu concerné. En particulier, $k_i(J)$ mesure le degré avec lequel l’agent $i$ s’identifie au groupe $J$. Mettre $k_i(J)$ à 0 correspond à un lien non-existant entre l’agent $i$ et le groupe $J$ (i.e., $i$ ne s’identifie pas à $J$), alors que mettre $k_i(J)$ à 1 signifie que l’agent $i$ est fortement lié au groupe $J$ (i.e., $i$ s’identifie fortement au groupe $J$). Il faut également noter que $k_i(\{i\})$ correspond à la mesure d’individualisme de l’agent $i$.

La fonction $k$ est également soumise aux contraintes suivantes:

1. $\textbf{C1}$ for every $i \in Agt$, $\sum_{J \in \text{Group}(i)} k_i(J) = 1$
Selon la contrainte $C_1$, $k_i$ défini, pour l’individu $i$, une distribution de liens sociaux avec chaque groupe $J = \{G \in 2^{\text{Agt}} | i \in G\}$. Dans ce cas, un agent ne peut ni s’identifier totalement à plusieurs groupes, ni s’identifier à aucun groupe (dans le pire cas, l’agent $i$ est extrêmement individualiste, i.e., $k_i(\{i\}) = 1$). A travers la contrainte $C_2$, nous assumons que le lien social est restreint à être bilatéral, c’est à dire, le degré de lien social avec un groupe $J$ est le même pour chaque membre de $J$.

Nous montrons ensuite les différents avantages offerts par un tel modèle par rapport à la théorie de Bacharach, particulièrement dans le contexte d’interactions sociales où différents groupes concurrents peuvent coexister.

Plus précisément, l’étude que nous présentons ici se concentre sur le concept central de l’identification de groupe qui permet de dévoiler le point focal requis pour résoudre les problèmes de coordination (Schelling [1960]).

Cependant, nous argumentons qu’une telle identification binaire de groupe (i.e., soit l’individu s’identifie au groupe ou pas), selon la théorie de raisonnement en équipe de Bacharach, n’est pas suffisante pour modéliser les situations de la vie de tous les jours où des individus s’identifient avec un groupe particulier avec un certain degré: par exemple, un individu peut être partagé entre s’identifier à un groupe d’amis, et s’identifier à un groupe de membres de sa famille. En conséquence, nous montrons comment la théorie que nous proposons permet de traiter efficacement ce genre de dilemmes.

8.7 Conclusion

Dans cette thèse, nous fournissons des arguments intuitifs expliquant pourquoi les gens deviennent souvent du comportement optimal. Bien qu’il soit souvent dit qu’un tel comportement résulte d’irrationalité, nous défendons plutôt sa compatibilité avec un principe de rationalité limité. En effet, la plupart des êtres humains sont incapables de raisonner correctement à propos de connaissances et croyances interactives complexes (e.g., à propos de ce que je sais que tu sais que je sais que tu sais que …). De la même manière, les gens ont également des mémoires limitées, ce qui les empêche de se souvenir parfaitement de ce qui s’est passé précédemment et d’utiliser cette information pour prédire ce qui se passera ensuite. Comme autre composante principale qui dirige les décisions rationnelles dans les interactions sociales, nous considérons également les relations sociales existantes entre les individus qui jouent un rôle majeur pour favoriser la coopération.

Cependant, bien que l’étude présentée dans cette thèse permet de clarifier le concept de rationalité sociale, il est clair qu’elle représente seulement un prémisse à
de multiple directions de recherche future qui méritent une attention particulière. En voici quelques exemples:

- Etendre l’étude logique présentée dans cette thèse afin de raisonner sur des événements passés, présents et futures, ce qui permettrait de raisonner sur des situations contrefactuelles. Une telle logique semble particulièrement intéressante pour une analyse détaillée des jeux extensifs avec information imparfaite ou information incomplète.

- Fournir une analyse logique de raisonnement normatif: l’étude expérimentale présenté dans cette thèse indique que les gens ne raisonnent pas stratégiquement, mais tendent à suivre ce qu’ils croient être la norme, c’est à dire, ce que tout le monde devrait faire dans une situation déterminée. La logique déontique semble alors représenter un outil adéquat pour répondre à cette problématique.

- Améliorer le modèle de liens sociaux présenté dans cette thèse qui considère seulement les relations avec des groupes (à travers le niveau avec lequel chaque individu s’identifie avec chaque groupe). En effet, un groupe est composé d’agents qui partagent des liens individuels les uns avec les autres. Cependant, exprimer les fonctions ci-dessus $k_i$ (pour chaque agent $i$) strictement en terme de relations individuelles n’apparait pas comme une tâche facile: un individu peut effectivement être très proche individuellement de chaque membre du groupe sans être proche socialement du groupe lui-même. Plus généralement, l’objectif de ce raffinement du modèle de liens sociaux est de fournir une analyse économique de l’impact de l’identification de groupe sur les interactions stratégiques au sein des réseaux sociaux. Une telle étude inclue également une analyse comparative avec les modèles économiques existant dans la littérature sur les réseaux sociaux (voir, e.g., Bramoullé et al. [2011]; Galeotti et al. [2010]; Goyal [2009]; Jackson [2005, 2007]).

- Analyser les liens sociaux et le raisonnement collectif dans le contexte d’interactions séquentielles. Par exemple, la relation entre trahison, punitions et liens sociaux représente un problème de recherche particulièrement pertinent: intuitivement, plus fort est la relation sociale entre deux individus, plus fort est l’impact d’une trahison de l’un envers l’autre, ce qui entraînera une punition plus importante de l’individu trahi envers l’acteur responsable (i.e., le traître). Plus généralement, ce genre d’étude suggère l’analyse du caractère dynamique des liens sociaux: en effet, il est raisonnable d’assumer qu’un comportement coopératif permet de renforcer une relation entre individus alors qu’un comportement individualiste l’affaibli. Des études expérimentales semblent alors requises afin d’examiner la création et l’évolution des
liens sociaux dans le contexte de situations interactives dynamiques (i.e., dans des jeux séquentiel ou des jeux répétés).

- Étudier expérimentalement les différences entre les prédictions de notre modèle de liens sociaux et celles de la théorie de raisonnement par équipe de Bacharach (e.g., voir le jeu de dictateur de la table 8.1).

- Comparer le modèle de liens sociaux présenté dans cette thèse avec le modèle d’Homo Moralis présenté dans Alger and Weibull [2012] qui suggère que le degré de moralité d’un individu dépend de la probabilité d’interagir avec des individus qui partagent les mêmes préférences.

- Clarifier ce qui défini une solution juste pour un groupe. Dans cette thèse, nous considérons deux approches pour calculer le profit collectif: soit les individus suivent le critère de justice de Rawls dont l’objectif est de maximiser la personne dans la situation la moins favorable, soit les individus suivent l’utilitarisme classique dont l’objectif est de maximiser l’efficacité globale du groupe. Cependant, laquelle de ces méthodes faut-il suivre reste une question ouverte dont la réponse dépend probablement du contexte. Afin d’examiner les hypothèses plausibles qui pourraient clarifier quelle fonction de profit collectif devrait être utilisée dans quel contexte (voir, e.g., Binmore [1998, 2005]), l’utilisation de simulations multi-agents apparaît alors comme un choix évident.
Appendix A

Proofs in modal logic ELEG

To make the proofs of Lemmas and Theorems more readable, we use the following abbreviation:
\[
\text{AllRat} \overset{\text{def}}{=} \bigwedge_{i \in \text{Agt}} \text{Rat}_i
\]

A.1 Proof of Lemmas

We first provide the proof of the following Lemmas that are necessary to later prove Theorems.

1. \( \vdash_{\text{ELEG}} [K_i] \text{Rat}_i \rightarrow \text{Rat}_i \)

2. \( \vdash_{\text{ELEG}} \text{Rat}_i \rightarrow [K_i] \text{Rat}_i \)

3. \( \vdash_{\text{ELEG}} [\text{CK}_i^{n+1}] \text{AX}^n \text{AllRat} \rightarrow [\text{CK}_i^n] \text{AX}^n \text{AllRat} \)

4. \( \vdash_{\text{ELEG}} \text{Depth}^n \rightarrow \text{AXDepth}^n \)

5. \( \vdash_{\text{ELEG}} \text{GenPos}^{n+1} \rightarrow \text{AXGenPos}^n \)

6. \( \vdash_{\text{ELEG}} (\text{Depth}^n \land \text{GenPos}^n \land k_i \land h_j) \rightarrow \Box (k_i \leftrightarrow h_j) \)

7. \( \vdash_{\text{ELEG}} (\text{Depth}^n \land \text{GenPos}^n \land k_i \land \text{BI}^n) \rightarrow \Box (\text{BI}^n \rightarrow k_i) \)
A.1.1 Syntactic proof of lemma A.1

We prove the following:

Lemma A.1 For every $i \in \text{Agt}$:

\[ \vdash_{\text{ELEG}} [K_i] \text{Rat}_i \rightarrow \text{Rat}_i \]

1. \[ \vdash_{\text{ELEG}} \text{end} \land \text{turn}_i \rightarrow \Box (\text{end} \land \text{turn}_i) \]
   by Axioms EndVert and TurnStr;

2. \[ \vdash_{\text{ELEG}} \neg \text{end} \land \text{turn}_i \rightarrow \Box (\neg \text{end} \land \text{turn}_i) \]
   by Axioms EndVert and TurnStr, and Axiom 5 for $\Box$;

3. \[ \vdash_{\text{ELEG}} \text{end} \land \text{turn}_i \land k_i \rightarrow \bigvee_{\alpha \in \text{Act}} \alpha_i \land k_i \]
   by Axioms OneAct;

4. \[ \vdash_{\text{ELEG}} \text{end} \land \text{turn}_i \land k_i \rightarrow [K_i]k_i \]
   by 3, Axioms EndAct, PerfectInfo, and Aware, and Axiom $K$ for $[K_i]$;

5. \[ \vdash_{\text{ELEG}} [K_i] \text{Rat}_i \rightarrow [K_i] \text{Rat}_i \land [K_i] \text{Rat}_i \neg \text{end} \]
   by the definitions of $\text{Rat}_i$ and boolean principles;

6. \[ \vdash_{\text{ELEG}} [K_i] \text{Rat}_i \neg \text{end} \rightarrow ((\text{end} \land \text{turn}_i) \rightarrow [K_i] \bigvee_{k \in I} (k_i \land \Box (V_{h \in I, h \leq k} h_i))) \]
   by 1, the definitions of $\text{Rat}_i \neg \text{end}$, Axiom PerfectInfo, and Axioms $T$ and $K$ for $[K_i]$ (or Axioms $D$ and $K$ if $[K_i]$ is $KD45$ modal operator), and Axiom 5 for $\Box$;

7. \[ \vdash_{\text{ELEG}} [K_i] \text{Rat}_i \neg \text{end} \rightarrow ((\text{end} \land \text{turn}_i) \rightarrow \bigvee_{k \prime \in I} k'_i \land [K_i] \bigvee_{k \in I} (k_i \land \Box (V_{h \in I, h \leq k} h_i))) \]
   by 6 and Axiom CompletePref;
8. \( \vdash_{ELEG} [K_i] \text{Rat}_{i}^{\text{end}} \) \\
\( \rightarrow ((\text{end} \land \text{turn}_i) \rightarrow \forall k \in I \ k_i \land [K_i](k_i \land \Box(\forall h \in I : h \leq k \ h_i)) \) \\
by 7 and 4, and Axiom **SinglePref**;

9. \( \vdash_{ELEG} [K_i] \text{Rat}_{i}^{\text{end}} \) \\
\( \rightarrow ((\text{end} \land \text{turn}_i) \rightarrow \forall k \in I \ k_i \land \Box(\forall h \in I : h \leq k \ h_i) \) \\
by 8, Axiom **PerfectInfo**, Axiom \( T \) for \([K_i]\) (or Axiom \( D \) if \([K_i]\) is \( KD45 \) modal operator), and Axiom 5 for \( \Box \);

10. \( \vdash_{ELEG} [K_i] \text{Rat}_{i}^{\text{end}} \) \\
\( \rightarrow \text{Rat}_{i}^{\text{end}} \) \\
by 9 and the definition of \( \text{Rat}_{i}^{\text{end}} \);

11. \( \vdash_{ELEG} [K_i] \text{Rat}_{i}^{\text{end}} \) \\
\( \rightarrow (\neg\text{end} \land \text{turn}_i) \rightarrow [K_i] \forall k \in I \ (k_i \land \mathsf{AX}(\forall h \in I : h \leq k \ (K_i) h_i)) \) \\
by 2, the definitions of \( \text{Rat}_{i}^{\text{end}} \), Axiom **PerfectInfo**, and Axioms \( T \) and \( K \) for \([K_i]\) (or Axioms \( D \) and \( K \) if \([K_i]\) is \( KD45 \) modal operator), and Axiom 5 for \( \Box \);

12. \( \vdash_{ELEG} [K_i] \text{Rat}_{i}^{\text{end}} \) \\
\( \rightarrow (\neg\text{end} \land \text{turn}_i) \rightarrow \forall k \in I \ (k_i \land \mathsf{AX}(\forall h \in I : h \leq k \ (K_i) h_i)) \) \\
by 11 and boolean principles;

13. \( \vdash_{ELEG} [K_i] \text{Rat}_{i}^{\text{end}} \) \\
\( \rightarrow (\neg\text{end} \land \text{turn}_i) \rightarrow \forall k \in I \ (k_i \land \mathsf{AX}(\forall h \in I : h \leq k \ (K_i) h_i)) \) \\
by 12 and Axiom 4 for \([K_i]\);

14. \( \vdash_{ELEG} [K_i] \text{Rat}_{i}^{\text{end}} \) \\
\( \rightarrow \text{Rat}_{i}^{\text{end}} \) \\
by 13 and the definition of \( \text{Rat}_{i}^{\text{end}} \);

15. \( \vdash_{ELEG} [K_i] \text{Rat}_{i} \) \\
\( \rightarrow \text{Rat}_{i}^{\text{end}} \land \text{Rat}_{i}^{\text{end}} \) \\
by 5, 10, and 14;
A.1.2 Syntactic proof of lemma A.2

We prove the following:

**Lemma A.2** For every \(i \in \text{Agt}\):

\[
\vdash_{\text{ELEG}} \text{Rat}_i \rightarrow [K_i] \text{Rat}_i
\]

1. \(\vdash_{\text{ELEG}} \text{end} \land \text{turn}_i\)
   \(\rightarrow \Box (\text{end} \land \text{turn}_i)\)
   by Axioms \text{EndVert} and \text{TurnStr};

2. \(\vdash_{\text{ELEG}} \neg \text{end} \land \text{turn}_i\)
   \(\rightarrow \Box (\neg \text{end} \land \text{turn}_i)\)
   by Axioms \text{EndVert} and \text{TurnStr}, and Axiom 5 for \(\Box\);

3. \(\vdash_{\text{ELEG}} \text{end} \land \text{turn}_i \land k_i\)
   \(\rightarrow \forall \alpha \in \text{Act} \alpha_i \land k_i\)
   by Axioms \text{OneAct};

4. \(\vdash_{\text{ELEG}} \text{end} \land \text{turn}_i \land k_i\)
   \(\rightarrow [K_i] k_i\)
   by 3, Axioms \text{EndAct}, \text{PerfectInfo}, and \text{Aware}, and Axiom \(K\) for \([K_i]\);

5. \(\vdash_{\text{ELEG}} \text{Rat}_i\)
   \(\rightarrow \text{Rat}_i^{\text{end}} \land \text{Rat}_i^{\neg \text{end}}\)
   by the definition of \(\text{Rat}_i\);

6. \(\vdash_{\text{ELEG}} \text{Rat}_i^{\text{end}}\)
   \(\rightarrow ((\text{end} \land \text{turn}_i) \rightarrow \forall k \in I (k_i \land \Box (V_{h \in I} h_i \leq k h_i)))\)
   by the definition of \(\text{Rat}_i^{\text{end}}\);
7. \[ \vdash_{\text{ELEG}} \text{Rat}^\text{end}_i \]
\[ \rightarrow ((\text{end} \land \text{turn}_i) \rightarrow \forall_{k \in I}(K_i)(k_i \land \Box(\forall_{h \in I : h \leq k}(h_i))) \]
by 6 and 4, Axiom \textbf{PerfectInfo}, and Axiom 4 for \( \Box \);

8. \[ \vdash_{\text{ELEG}} \text{Rat}^\text{end}_i \]
\[ \rightarrow [K_i]((\text{end} \land \text{turn}_i) \rightarrow \forall_{k \in I}(K_i)(k_i \land \Box(\forall_{h \in I : h \leq k}(h_i))) \]
by 7 and 1, Axiom \textbf{PerfectInfo}, and boolean principles;

9. \[ \vdash_{\text{ELEG}} \text{Rat}^\text{end}_i \]
\[ \rightarrow [K_i]\text{Rat}^\text{end}_i \]
by 8 and the definition of \( \text{Rat}^\text{end}_i \);

10. \[ \vdash_{\text{ELEG}} \text{Rat}^{-\text{end}}_i \]
\[ \rightarrow ((\neg \text{end} \land \text{turn}_i) \rightarrow \forall_{k \in I}(K_i)(k_i \land \Box(\forall_{h \in I : h \leq k}(K_i)h_i))) \]
by the definition of \( \text{Rat}^{-\text{end}}_i \);

11. \[ \vdash_{\text{ELEG}} \text{Rat}^{-\text{end}}_i \]
\[ \rightarrow ((\neg \text{end} \land \text{turn}_i) \rightarrow [K_i] \forall_{k \in I}(K_i)(k_i \land \Box(\forall_{h \in I : h \leq k}(K_i)h_i))) \]
by 10, Axiom 5 for \( [K_i] \), and boolean principles;

12. \[ \vdash_{\text{ELEG}} \text{Rat}^{-\text{end}}_i \]
\[ \rightarrow [K_i]((\neg \text{end} \land \text{turn}_i) \rightarrow \forall_{k \in I}(K_i)(k_i \land \Box(\forall_{h \in I : h \leq k}(K_i)h_i))) \]
by 11 and 2, Axiom \textbf{PerfectInfo}, and boolean principles;

13. \[ \vdash_{\text{ELEG}} \text{Rat}^{-\text{end}}_i \]
\[ \rightarrow [K_i]\text{Rat}^{-\text{end}}_i \]
by 12 and the definition of \( \text{Rat}^{-\text{end}}_i \);

14. \[ \vdash_{\text{ELEG}} \text{Rat}_i \]
\[ \rightarrow [K_i]\text{Rat}^{-\text{end}}_i \land [K_i]\text{Rat}^{-\text{end}}_i \]
by 5, 9, and 13;

15. \[ \vdash_{\text{ELEG}} \text{Rat}_i \]
\[ \rightarrow [K_i]\text{Rat}_i \]
by 14 and boolean principles;
A.1.3 Syntactic proof of lemma A.3

We prove the following:

Lemma A.3 For every \( n \in \mathbb{N} \):

\[
\vdash^{ELEG} [\text{CK}_{Agt}^{n+1}]AX^n \text{AllRat} \rightarrow [\text{CK}_{Agt}^{n}]AX^n \text{AllRat}
\]

Basic case \((n = 0)\):

1. \( \vdash^{ELEG} [\text{CK}_{Agt}^{1}]\text{AllRat} \rightarrow [\text{EK}_{Agt}]\text{AllRat} \)
   by the definition of \([\text{CK}_{Agt}^{1}]\); 

2. \( \vdash^{ELEG} [\text{CK}_{Agt}^{1}]\text{AllRat} \rightarrow \bigwedge_{i \in \text{Agt}} [K_i] \text{Rat}_i \)
   by 1 and the definitions of \([\text{EK}_{Agt}^{1}]\) and \(\text{AllRat}\); 

3. \( \vdash^{ELEG} [\text{CK}_{Agt}^{1}]\text{AllRat} \rightarrow \bigvee_{i \in \text{Agt}} \text{Rat}_i \)
   by 2 and Lemma A.1; 

4. \( \vdash^{ELEG} [\text{CK}_{Agt}^{1}]\text{AllRat} \rightarrow \text{AllRat} \)
   by 3 and the definition of \(\text{AllRat}\) (i.e. \([\text{CK}_{Agt}^{0}]\text{AllRat}\));

General case (for \( n > 0 \)):

1. \( \vdash^{ELEG} [\text{CK}_{Agt}^{n+1}]AX^n \text{AllRat} \rightarrow \bigwedge_{1 \leq k \leq n+1} [\text{EK}_{Agt}^{k}]AX^n \text{AllRat} \)
   by the definition of \([\text{CK}_{Agt}^{n+1}]\); 

2. \( \vdash^{ELEG} [\text{CK}_{Agt}^{n+1}]AX^n \text{AllRat} \rightarrow \bigwedge_{1 \leq k \leq n} [\text{EK}_{Agt}^{k}]AX^n \text{AllRat} \land [\text{EK}_{Agt}^{n+1}]AX^n \text{AllRat} \)
   by 1 and boolean principles; 

3. \( \vdash^{ELEG} [\text{CK}_{Agt}^{n+1}]AX^n \text{AllRat} \rightarrow [\text{CK}_{Agt}^{n}]AX^n \text{AllRat} \)
   by 2 and the definition of \([\text{CK}_{Agt}^{n}]\);
A.1.4 Syntactic proof of lemma A.4

We prove the following:

Lemma A.4 For every $n \in \mathbb{N}$:

$$\vdash_{\text{ELEG}} \text{Depth}^{n+1} \rightarrow \text{AXDepth}^n$$

1. $\vdash_{\text{ELEG}} \text{Depth}^{n+1}$
   $\rightarrow \Box^n(\Box X)^{n}\text{end}$
   by definition of $\text{Depth}^{n+1}$;

2. $\vdash_{\text{ELEG}} \text{Depth}^{n+1}$
   $\rightarrow \text{AX}(\Box X)^{n}\text{end}$
   by 1, Axioms 4 and $T$ for $\Box$ and Axiom $\text{NxtVert}$;

3. $\vdash_{\text{ELEG}} \text{Depth}^{n+1}$
   $\rightarrow \text{AXDepth}^n$
   by 2 and definition of $\text{Depth}^n$;

A.1.5 Syntactic proof of lemma A.5

We prove the following:

Lemma A.5 For every $n \in \mathbb{N}$:

$$\vdash_{\text{ELEG}} \text{GenPos}^{n+1} \rightarrow \text{AXGenPos}^n$$

1. $\vdash_{\text{ELEG}} \text{GenPos}^{n+1}$
   $\rightarrow \bigwedge_{0 \leq k \leq n+1} \bigwedge_{i \in \text{Agt}, \epsilon \in \text{Seq}^k} \text{AX}^{\leq n+1}\Box \left((k_i \wedge (\epsilon)\text{end}) \rightarrow \Box (\epsilon)\text{end} \leftrightarrow k_i\right)$
   by definition of $\text{GenPos}^{n+1}$;

2. $\vdash_{\text{ELEG}} \text{GenPos}^{n+1}$
   $\rightarrow \text{AX} \bigwedge_{0 \leq k \leq n+1} \bigwedge_{i \in \text{Agt}, \epsilon \in \text{Seq}^k} \text{AX}^{\leq n} \Box \left((k_i \wedge (\epsilon)\text{end}) \rightarrow \Box (\epsilon)\text{end} \leftrightarrow k_i\right)$
   by 1, Theorem $\vdash_{\text{ELEG}} \text{AX}^{n+1}\varphi \rightarrow \text{AXAX}^{\leq n}\varphi$, and boolean principles;

3. $\vdash_{\text{ELEG}} \text{GenPos}^{n+1} \rightarrow \text{AXGenPos}^n$
   by 2 and the definition of $\text{GenPos}^n$. 

204
A.1.6 Syntactic proof of lemma A.6

We prove the following inductively:

**Lemma A.6** For every $n \in \mathbb{N}$, $i, j \in \text{Agt}$:

$$\vdash_{ELEG} (\text{Depth}^n \land \text{GenPos}^n \land k_i \land h_j) \rightarrow \Box (k_i \leftrightarrow h_j)$$

**Basic case $n=0$:**

Here, we prove

$$\vdash_{ELEG} \text{end} \land \text{GenPos}^0 \land k_i \land h_j \rightarrow \Box (k_i \leftrightarrow h_j)$$

1. $\vdash_{ELEG} \text{end} \land \text{GenPos}^0 \land k_i \land h_j$
   $\rightarrow \forall \alpha \in \text{Act} \alpha \land \Box((\alpha \land \text{end}) \leftrightarrow k_i) \land \Box((\alpha \land \text{end}) \leftrightarrow h_j)$
   by definition of GenPos$^0$, Axiom OneAct, and Axiom T for $\Box$;

2. $\vdash_{ELEG} \text{end} \land \text{GenPos}^0 \land k_i \land h_j$
   $\rightarrow \Box (k_i \leftrightarrow h_j)$
   by 1 and boolean principles;

**Inductive case:**

Let $n \in \mathbb{N}$ and let us prove that if the theorem is true for all $k \leq n$, then it is true for $n + 1$.

1. $\vdash_{ELEG} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_i \land h_j$
   $\rightarrow k_i \land h_j \land X (k_i \land h_j \land \text{Depth}^n \land \text{GenPos}^n)$
   by Axiom TimePref, and Lemmas A.4 and A.5;

2. $\vdash_{ELEG} \text{Depth}^n \land \text{GenPos}^n \land k_i \land h_j$
   $\rightarrow \Box (k_i \leftrightarrow h_j)$
   by induction;

3. $\vdash_{ELEG} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_i \land h_j$
   $\rightarrow \forall \alpha \in \text{Act} \alpha \land k_i \land X \Box (k_i \leftrightarrow h_j)$
   by 1 and 2, and Axiom OneAct;
4. \( \vdash_{ELEG} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_i \land h_j \)
   \( \rightarrow \bigvee_{\alpha \in \text{Act}} \alpha \land k_i \land \Box(\alpha \rightarrow X(k_i \leftrightarrow h_j)) \)
   by 3, and Axiom \textbf{StrAct};

5. \( \vdash_{ELEG} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_i \land h_j \)
   \( \rightarrow \bigvee_{\alpha \in \text{Act}, \epsilon \in \text{Seq}^n} \alpha \land X(\epsilon)_{\text{end}} \land k_i \land \Box(\alpha \rightarrow (k_i \leftrightarrow h_j)) \)
   by 4, Axiom \textbf{TimePref} and the definition of \text{Depth}^{n+1};

6. \( \vdash_{ELEG} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_i \land h_j \)
   \( \rightarrow \bigvee_{\alpha \in \text{Act}, \epsilon \in \text{Seq}^n} \Box(k_i \leftrightarrow (\alpha \land X(\epsilon)_{\text{end}})) \land \Box(\alpha \rightarrow (k_i \leftrightarrow h_j)) \)
   by 5 and the definition of \text{GenPos}^{n+1};

7. \( \vdash_{ELEG} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_i \land h_j \)
   \( \rightarrow \bigvee_{\alpha \in \text{Act}} \Box(\alpha \rightarrow (k_i \leftrightarrow h_j)) \)
   by 6 and boolean principles;

8. \( \vdash_{ELEG} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_i \land h_j \)
   \( \rightarrow \bigvee_{\alpha \in \text{Act}} \Box(\alpha \leftrightarrow (k_i \leftrightarrow h_j)) \)
   by 7, and boolean principles;

9. \( \vdash_{ELEG} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_i \land h_j \)
   \( \rightarrow \Box(k_i \leftrightarrow h_j) \)
   by 8, and boolean principles;

\textbf{A.1.7 Syntactic proof of lemma A.7}

We prove the following:

\textbf{Lemma A.7} For every \( n \in \mathbb{N}, i, j \in \text{Agt} \):

\( \vdash_{ELEG} (\text{Depth}^n \land \text{GenPos}^n \land k_j \land \text{Bl}^n) \rightarrow \Box(\text{Bl}^n \rightarrow k_j) \)

\textbf{Basic case (}\( n = 0 \):\)

Here, we prove:

\( \vdash_{ELEG} (\text{end} \land \text{GenPos}^0 \land k_j \land \text{Bl}^0) \rightarrow \Box(\text{Bl}^0 \rightarrow k_j) \)
1. \( \vdash_{\text{ELEG}} (\text{end} \land \text{GenPos}^0 \land k \land \text{Bl}^0) \)
   \[ \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k_i' \land \text{Bl}^0 \land \Diamond (\text{turn}_i \land k''_i \land \text{Bl}^0) \]
   by Axioms \text{TurnTaking}, \text{TurnStr} and \text{CompletePref} and Axiom \( T \) for \( \Box \);

2. \( \vdash_{\text{ELEG}} (\text{end} \land \text{GenPos}^0 \land k \land \text{Bl}^0) \)
   \[ \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k_i' \land \Diamond (\text{turn}_i \land k''_i) \land \Box (V_{h \in I : h \leq k', h \leq k''} h_i) \]
   by 1 and the definition of \( \Box \);

3. \( \vdash_{\text{ELEG}} (\text{end} \land \text{GenPos}^0 \land k \land \text{Bl}^0) \)
   \[ \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k_i' \land \Diamond (\text{turn}_i \land k''_i) \land \Box (V_{h \in I : h \leq k', h \leq k''} h_i) \]
   by 2, Axiom 5 for \( \Box \), and boolean principles;

4. \( \vdash_{\text{ELEG}} (\text{end} \land \text{GenPos}^0 \land k \land \text{Bl}^0) \)
   \[ \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k_i' \land V_{h \in I : h \leq k', h \leq k''} h_i \land \Diamond (\text{turn}_i \land k''_i \land V_{h \in I : h \leq k', h \leq k''} h_i) \]
   by 5, Axiom \( K \) and \( T \) for \( \Box \), and boolean principles;

5. \( \vdash_{\text{ELEG}} (\text{end} \land \text{GenPos}^0 \land k \land \text{Bl}^0) \)
   \[ \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k_i' \land V_{h \in I : h \leq k', h \leq k''} h_i \land \Diamond (\text{turn}_i \land k''_i \land \text{Bl}^0) \]
   by 7, Axiom \text{SinglePref}, and boolean principles;

6. \( \vdash_{\text{ELEG}} (\text{end} \land \text{GenPos}^0 \land k \land \text{Bl}^0) \)
   \[ \rightarrow (V_{i \in \text{Agt}, k', k'' \in I : k' \neq k''} \text{turn}_i \land k_i' \land \text{Bl}^0 \land \Diamond (\text{turn}_i \land k''_i \land \text{Bl}^0)) \rightarrow \bot \]
   by 8, and boolean principles;

7. \( \vdash_{\text{ELEG}} (\text{end} \land \text{GenPos}^0 \land k \land \text{Bl}^0) \)
   \[ \rightarrow V_{i \in \text{Agt}, k' \in I} \text{turn}_i \land k_i' \land \Box (\text{Bl}^0 \rightarrow k'_i) \]
   by 9 and boolean principles;

8. \( \vdash_{\text{ELEG}} (\text{end} \land \text{GenPos}^0 \land k \land \text{Bl}^0) \)
   \[ \rightarrow \Box (\text{Bl}^0 \rightarrow k_j) \]
   by 10 and Lemma A.6;

\text{Inductive case:}

Let \( n \in \mathbb{N} \) and let us prove that if the theorem is true for all \( k \leq n \), then it is true for \( n + 1 \).
1. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{BI}^{n+1} \)
   \( \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k'_i \land \text{BI}^{n+1} \land \Diamond (\text{turn}_i \land k''_i \land \text{BI}^{n+1}) \)
   by Axioms \text{TurnTaking}, \text{TurnStr} and \text{CompletePref} and Axiom \text{T} for \( \Box \);

2. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{BI}^{n+1} \)
   \( \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k'_i \land \text{AX(\text{BI}^{n} \land V_{h \in I : h \leq k'} \text{h}_i}) \land \Diamond (\text{turn}_i \land k''_i \land \text{AX(\text{BI}^{n} \land V_{h \in I : h \leq k''} \text{h}_i})) \)
   by 1 and the definition of \( \text{BI}^{n+1} \);

3. \( \vdash_{\text{ELEG}} \text{Depth}^{n} \land \text{GenPos}^{n} \land k_j \land \text{BI}^{n} \)
   \( \rightarrow \Box (\text{BI}^{n} \rightarrow k_j) \)
   by induction;

4. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{BI}^{n+1} \)
   \( \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k'_i \land \text{AX(\text{BI}^{n} \land V_{h \in I : h \leq k'} \text{h}_i}) \land \Diamond (\text{turn}_i \land k''_i \land \text{AX(\text{BI}^{n} \land V_{h \in I : h \leq k''} \text{h}_i})) \)
   by 2 and 3, Lemmas A.4 and A.5, Axiom \text{Perm}_{\Box, \text{AX}}, and Axiom 4 for \( \Box \);

5. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{BI}^{n+1} \)
   \( \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k'_i \land \text{AXBI}^{n} \land \Diamond (\text{turn}_i \land k''_i \land \text{AXBI}^{n}) \land \Box (\text{BI}^{n} \rightarrow V_{h \in I : h \leq k'} \text{h}_i)) \)
   by 4 and 2, Axiom \text{Perm}_{\Box, \text{AX}}, Axiom 5 for \( \Box \);

6. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{BI}^{n+1} \)
   \( \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k'_i \land \text{AX(V}_{h \in I : h \leq k'} \text{h}_i) \land \Diamond (\text{turn}_i \land k''_i \land \text{AX(V}_{h \in I : h \leq k''} \text{h}_i)) \)
   by 5, Axiom \text{T} for \( \Box \), and boolean principles;

7. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{BI}^{n+1} \)
   \( \rightarrow V_{i \in \text{Agt}, k', k'' \in I} \text{turn}_i \land k'_i \land V_{h \in I : h \leq k'} \text{h}_i \land \Diamond (\text{turn}_i \land k''_i \land V_{h \in I : h \leq k''} \text{h}_i) \)
   by 6, the definition of \text{Depth}^{n+1}, Axioms \text{TimeVert} and \text{TimePref};

8. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{BI}^{n+1} \)
   \( \rightarrow V_{i \in \text{Agt}, k', k'' \in I : k' \leq k'', k'' \leq k'} \text{turn}_i \land k'_i \land \text{BI}^{n+1} \land \Diamond (\text{turn}_i \land k''_i \land \text{BI}^{n+1}) \)
   by 7, Axiom \text{SinglePref}, and boolean principles;
9. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{Bl}^{n+1} \)
\( \rightarrow (\forall \{i \in \text{Agt}, k', k'' \mid k' \neq k''\} \text{ turn}_i \land k'_i \land \text{Bl}^{n+1} \land \diamond (\text{turn}_i \land k''_i \land \text{Bl}^{n+1})) \rightarrow \bot \)
by 8, and boolean principles;

10. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{Bl}^{n+1} \)
\( \rightarrow \forall \{i \in \text{Agt}, k' \mid k' \in I\} \text{ turn}_i \land k'_i \land \text{Bl}^{n+1} \land \Box(\text{Bl}^{n+1} \rightarrow k'_i) \)
by 9 and boolean principles;

11. \( \vdash_{\text{ELEG}} \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \land k_j \land \text{Bl}^{n+1} \)
\( \rightarrow \Box(\text{Bl}^{n+1} \rightarrow k_j) \)
by 10 and Lemma A.6;

### A.2 Proofs of Theorems

We here provide the proof of the following Theorems.

1. \( \vdash_{\text{ELEG}} \text{Nash}^{n+1} \rightarrow \text{XNash}^n \)

2. \( \vdash_{\text{ELEG}} \text{Bl}^n \leftrightarrow \land_{0 \leq m \leq n} \text{AX}^m \text{Nash}^{n-m} \)

3. \( \vdash_{\text{ELEG}} \text{SRat}_i^n \leftrightarrow \lceil K_i \rceil \text{SRat}_i^n \)

4. \( \vdash_{\text{ELEG}} (\lceil \text{CK}_{\text{Agt}}^n \rceil \land \land_{i \in \text{Agt}} \text{SRat}_i^n \land \text{Depth}^n \land \text{GenPos}^n) \rightarrow \text{Bl}^n \)

### A.2.1 Syntactic proof of theorem 3.1

We prove the following (for every \( n \in \mathbb{N} \)):

1. \( \vdash_{\text{ELEG}} \text{Nash}^{n+1} \rightarrow \text{XNash}^n \)

\( \vdash_{\text{ELEG}} \text{Nash}^{n+1} \rightarrow \neg \text{end} \land \land_{i \in \text{Agt}} \lor_{k \in I} k_i \land \text{BR}^{n+1}(i, k) \)
by the definition of \( \text{Nash}^{n+1} \);
3. $\vdash_{\text{ELEG}} \text{Nash}^{n+1} \rightarrow X(\bigwedge_{i \in \text{Agt}} \bigvee_{k \in I} \text{BR}^n(i, k))$
   by 2 and Axiom \text{TimeVert};

4. $\vdash_{\text{ELEG}} \text{Nash}^{n+1} \rightarrow X\text{Nash}^n$
   by 3 and the definition of \text{Nash}^n;

\textbf{A.2.2 Syntactic proof of theorem 3.2}

We prove the following inductively (for every $n \in \mathbb{N}$):

$$\vdash_{\text{ELEG}} \text{Bl}^n \leftrightarrow \bigwedge_{0 \leq m \leq n} \text{AX}^m \text{Nash}^{n-m}$$

\textbf{Basic case $n=0$:}

Here, we prove $\vdash_{\text{ELEG}} \text{Bl}^0 \leftrightarrow \text{Nash}^0$.

1. $\vdash_{\text{ELEG}} \text{Bl}^0 \leftrightarrow \text{end} \land \bigvee_{i \in \text{Agt}, k \in I} \text{turn}_i \land k_i \land \Box (\forall_{h \in I \land h \leq k_i} h_i)$
   by the definition of \text{Bl}^0;

2. $\vdash_{\text{ELEG}} \text{Bl}^0 \leftrightarrow \text{end} \land \bigvee_{i \in \text{Agt}, k \in I} \text{turn}_i \land k_i \land (\text{turn}_i \rightarrow \Box (\forall_{h \in I \land h \leq k_i} h_i))$
   by 1 and boolean principle;

3. $\vdash_{\text{ELEG}} \text{Bl}^0 \leftrightarrow \text{end} \land \bigwedge_{i \in \text{Agt}} \bigvee_{k \in I} k_i \land (\text{turn}_i \land \Box (\forall_{h \in I \land h \leq k_i} h_i))$
   by 2, Axioms \text{SingleTurn} and \text{CompletePref}, and boolean principles;

4. $\vdash_{\text{ELEG}} \text{Bl}^0 \leftrightarrow \bigwedge_{i \in \text{Agt}} \bigvee_{k \in I} k_i \land \text{BR}^0(i, k)$
   by 3 and the definition of \text{BR}^0(i, k);

5. $\vdash_{\text{ELEG}} \text{Bl}^0 \leftrightarrow \text{Nash}^0$
   by 4 and the definition of \text{Nash}^0;

\textbf{Inductive case:}

Let $n \in \mathbb{N}$ and let us prove that if the theorem is true for all $k \leq n$, then it is true for $n+1$.

1. $\vdash_{\text{ELEG}} \text{Bl}^{n+1} \leftrightarrow \neg \text{end} \land \bigvee_{i \in \text{Agt}, k \in I} \text{turn}_i \land k_i \land \text{AX} (\text{Bl}^n \land \bigvee_{h \in I \land h \leq k_i} h_i)$
   by the definition of \text{Bl}^{n+1};
2. $\vdash_{\text{ELEG}} \text{Bi}^n \leftrightarrow \bigwedge_{0 \leq m \leq n} \text{AX}^m \text{Nash}^{n-m}$
   by induction;

3. $\vdash_{\text{ELEG}} \text{Bi}^{n+1} \leftrightarrow$
   $\neg \text{end} \land \bigwedge_{i \in \text{Agt}, k \leq l} \text{turn}_i \land k_i \land \text{AX}(\bigwedge_{0 \leq m \leq n} \text{AX}^m \text{Nash}^{n-m} \land \forall h \in \ell, h \leq k \ h_i)$
   by 1 and 2;

4. $\vdash_{\text{ELEG}} \text{Bi}^{n+1} \leftrightarrow$
   $\neg \text{end} \land \bigwedge_{i \in \text{Agt}, k \leq l} \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \land \bigwedge_{1 \leq m \leq n+1} \text{AX}^m \text{Nash}^{n+1-m}$
   by 3, $K$ principles for $\text{AX}$, the definition of $\text{AX}^m$, and boolean principles;

5. $\vdash_{\text{ELEG}} \neg \text{end} \land \bigwedge_{i \in \text{Agt}, k \leq l} \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \leftrightarrow$
   $\forall i \in \text{Agt}, k \leq l \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \land (\text{turn}_i \rightarrow \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n))$
   by Axioms $\text{TimeVert}$ and $\text{EndVert}$;

6. $\vdash_{\text{ELEG}} \neg \text{end} \land \bigwedge_{i \in \text{Agt}, k \leq l} \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \leftrightarrow$
   $\forall i \in \text{Agt}, k \leq l \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \land (\text{turn}_i \rightarrow \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n))$
   by 5 and boolean principles;

7. $\vdash_{\text{ELEG}} \neg \text{end} \land \bigwedge_{i \in \text{Agt}, k \leq l} \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \leftrightarrow$
   $\bigwedge_{i \in \text{Agt}} \forall k \leq l \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \land (\text{turn}_i \rightarrow \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n))$
   by 6, Axioms $\text{SingleTurn}$ and $\text{CompletePref}$, and boolean principles;

8. $\vdash_{\text{ELEG}} \neg \text{end} \land \bigwedge_{i \in \text{Agt}, k \leq l} \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \leftrightarrow$
   $\bigwedge_{i \in \text{Agt}} \forall k \leq l \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{BR}^n(i, h)) \land (\text{turn}_i \rightarrow \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{BR}^n(i, h)))$
   by 7 and the definition of $\text{Nash}^n$;

9. $\vdash_{\text{ELEG}} (\forall h \in \ell, h \leq k \text{BR}^n(i, h)) \leftrightarrow \text{BR}^n(i, k)$
   by the definition of $\text{BR}^n(i, h)$;

10. $\vdash_{\text{ELEG}} \neg \text{end} \land \bigwedge_{i \in \text{Agt}, k \leq l} \text{turn}_i \land k_i \land \text{AX}(\forall h \in \ell, h \leq k \ h_i \land \text{Nash}^n) \leftrightarrow$
    $\bigwedge_{i \in \text{Agt}} \forall k \leq l \text{turn}_i \land k_i \land \text{XBR}^n(i, k) \land (\text{turn}_i \rightarrow \text{AXBR}^n(i, k))$
    by 8 and 9;
11. $\vdash_{\text{ELEG}} \text{¬end} \land \bigwedge_{i \in \text{Agt}, k \in I} \text{turn}_i \land k_i \land \text{AX}(\bigvee_{h \in I, h \leq_k} h_i \land \text{Nash}^n) \leftrightarrow \bigwedge_{i \in \text{Agt}} \bigvee_{k \in I} \text{BR}^{n+1}(i, k)$
   by 10 and the definition of $\text{BR}^{n+1}(i, k)$;

12. $\vdash_{\text{ELEG}} \text{¬end} \land \bigwedge_{i \in \text{Agt}, k \in I} \text{turn}_i \land k_i \land \text{AX}(\bigvee_{h \in I, h \leq_k} h_i \land \text{Nash}^n) \leftrightarrow \text{Nash}^{n+1}$
   by 11 and the definition of $\text{Nash}^{n+1}$;

13. $\vdash_{\text{ELEG}} \text{BI}^{n+1} \leftrightarrow (\text{Nash}^{n+1} \land \bigwedge_{1 \leq m \leq n+1} \text{AX}^m \text{Nash}^{n+1-m})$
   by 4 and 12;

14. $\vdash_{\text{ELEG}} \text{BI}^{n+1} \leftrightarrow \bigwedge_{0 \leq m \leq n+1} \text{AX}^m \text{Nash}^{n+1-m}$
   by 13, the definition of $\text{AX}^0$, and boolean principles;

A.2.3 Syntactic proof of theorem 3.3

We prove the following (for every $n \in \mathbb{N}, i \in \text{Agt}$):

$\vdash_{\text{ELEG}} \text{SRat}^n_i \leftrightarrow [K_i] \text{SRat}^n_i$

1. $\vdash_{\text{ELEG}} \text{Rat}_i \leftrightarrow [K_i] \text{Rat}_i$
   by Lemmas A.1 and A.2;

2. $\vdash_{\text{ELEG}} \text{AX}^{\leq n} \text{Rat}_i \leftrightarrow \text{AX}^{\leq n} [K_i] \text{Rat}_i$
   by 1 and Axiom $K$ for $\text{AX}$, and the definition of $\text{AX}^{\leq n}$;

3. $\vdash_{\text{ELEG}} \text{AX}^{\leq n} \text{Rat}_i \leftrightarrow [K_i] \text{AX}^{\leq n} \text{Rat}_i$
   by 2 and Axiom $\text{Perm}_{[K_i], \text{AX}}$;

4. $\vdash_{\text{ELEG}} \text{SRat}^n_i \leftrightarrow [K_i] \text{SRat}^n_i$
   by 3 and the definition of $\text{SRat}^n_i$;

A.2.4 Syntactic proof of theorem 3.4

We demonstrate Aumann’s theorem, which states the following (for every $n \in \mathbb{N}$):

$\vdash_{\text{ELEG}} ([\text{CK}^n_{\text{Agt}}] \bigwedge_{i \in \text{Agt}} \text{SRat}^n_i \land \text{Depth}^n \land \text{GenPos}^n) \rightarrow \text{BI}^n$
It is straightforward to show through boolean principles that:

\[ \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n} [\text{CK}_{Agt}^m]_n \text{AllRat} \rightarrow \bigwedge_{0 \leq m \leq n} [\text{CK}_{Agt}^m]_m \text{AllRat} ) \]

We therefore prove the following inductively:

\[ \vdash_{\text{ELEG}} ( \bigwedge_{0 \leq m \leq n} [\text{CK}_{Agt}^m]_m \text{AllRat} \land \text{Depth}^n \land \text{GenPos}^n ) \rightarrow \text{BI}^n \]

**Basic case \( n=0 \):**

Here, we prove \( \vdash_{\text{ELEG}} \text{AllRat} \land \text{end} \land \text{GenPos}^0 \rightarrow \text{BI}^0 \).

1. \( \vdash_{\text{ELEG}} (\text{AllRat} \land \text{end} \land \text{GenPos}^0) \)
   \[ \rightarrow \bigvee_{i \in \text{Agt}} (\text{turn}_i \land \text{Rat}_{i}^\text{end}) \]
   by the definition of \( \text{Rat}_{i}^\text{end} \), and Axiom TurnTaking;

2. \( \vdash_{\text{ELEG}} (\text{AllRat} \land \text{end} \land \text{GenPos}^0) \)
   \[ \rightarrow \bigvee_{i \in \text{Agt}} (\text{end} \land \text{turn}_i \land \bigvee_{k \in I} \text{Rat}_{k}^i \land \Box (\bigvee_{h \in I} \text{Rat}_{h}^i) \)
   by 1, and the definition of \( \text{Rat}_{i}^\text{end} \);

3. \( \vdash_{\text{ELEG}} (\text{AllRat} \land \text{end} \land \text{GenPos}^0) \rightarrow \text{BI}^0 \)
   by 2 and the definition of \( \text{BI}^0 \);

**Inductive case:**

Let \( n \in \mathbb{N} \) and let us prove that if the theorem is true for all \( k \leq n \), then it is true for \( n + 1 \).

1. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{Agt}^m]_m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1} ) \)
   \[ \rightarrow (\bigwedge_{0 \leq m \leq n} [\text{CK}_{Agt}^m]_m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1} ) \]
   by Theorem \( \vdash_{\text{ELEG}} \text{AX}^{n+1} \varphi \rightarrow \text{AXAX}^n \varphi \);

2. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{Agt}^m]_m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1} ) \)
   \[ \rightarrow \text{AX}(\bigwedge_{0 \leq m \leq n} [\text{CK}_{Agt}^m]_m \text{AllRat} \land \text{Depth}^n \land \text{GenPos}^n ) \]
   by 1, Lemmas A.4 and A.5, and Axiom Perm\text{[K]}_{\text{AX}} (or Perm\text{*[K]}_{\text{AX}});
3. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n+1} [CK^m_{Agl}] AX^m AllRat \land Depth^{n+1} \land GenPos^{n+1}) \)
   \( \rightarrow AX(\bigwedge_{0 \leq m \leq n} [CK^m_{Agt}] AX^m AllRat \land Depth^n \land GenPos^n) \)
   by 2 and Lemma A.3;

4. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n+1} [CK^m_{Agt}] AX^m AllRat \land Depth^{n+1} \land GenPos^{n+1}) \)
   \( \rightarrow [EK^1_{Agt}] AX(\bigwedge_{0 \leq m \leq n} [CK^m_{Agt}] AX^m AllRat \land Depth^n \land GenPos^n) \)
   by 1, Lemmas A.4 and A.5, and Axiom \text{Perm}_{[K_i],AX} \) (or \text{Perm}^*_{{[K_i],AX}});

5. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n} [CK^m_{Agt}] AX^m AllRat \land Depth^n \land GenPos^n) \rightarrow BI^n \)
   by induction;

6. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n+1} [CK^m_{Agt}] AX^m AllRat \land Depth^{n+1} \land GenPos^{n+1}) \)
   \( \rightarrow AXBI^n \land [EK^1_{Agt}] AXBI^n \)
   by 3, 4 and 5;

7. \( \vdash_{\text{ELEG}} GenPos^n \rightarrow [K_j] GenPos^n \)
   by Axioms \text{Perm}_{\Box AX} \) and \text{PerfectInfo}, and Axiom 4 for \( \Box \) and \([K_j]\);

8. \( \vdash_{\text{ELEG}} Depth^n \rightarrow [K_j] Depth^n \)
   by Axiom \text{PerfectInfo}, and Axiom 4 for \( \Box \);

9. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n+1} [CK^m_{Agt}] AX^m AllRat \land Depth^{n+1} \land GenPos^{n+1}) \)
   \( \rightarrow AX(V_{k \in I} k_i \land BI^n \land [K_j] BI^n \land Depth^n \land GenPos^n) \land [K_j] AX(V_{k \in I} k_i \land BI^n \land [K_j] BI^n \land Depth^n \land GenPos^n) \)
   by 6, 7, and 8, Lemmas A.4 and A.5, Axioms \text{CompletePref} \) and \text{Perm}_{[K_i],AX} \) (or \text{Perm}^*_{{[K_i],AX}}), and Axiom 4 for \([K_i]\);

10. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n+1} [CK^m_{Agt}] AX^m AllRat \land Depth^{n+1} \land GenPos^{n+1}) \)
    \( \rightarrow V_{i \in Agt, a \in Act} turn_t \land \alpha_i \land [K_i] \alpha_i \land X V_{k \in I} (BI^n \rightarrow k_i) \)
    by 9, Lemma A.7, and Axioms \text{TurnTaking, TimeVert, OneAct, and Aware};

11. \( \vdash_{\text{ELEG}} (\bigwedge_{0 \leq m \leq n+1} [CK^m_{Agt}] AX^m AllRat \land Depth^{n+1} \land GenPos^{n+1}) \)
    \( \rightarrow V_{i \in Agt, a \in Act} turn_t \land \alpha_i \land [K_i] \alpha_i \land V_{k \in I} (\alpha_i \rightarrow X(BI^n \rightarrow k_i)) \)
    by 10 and Axiom \text{StrAct};

214
12. \[\vdash \text{ELEG} \left(\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{m,\text{Ag}}] \text{AX}^m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1}\right)\]
   \[\to \forall i \in \text{Ag}, \text{turn}_i \land \text{XBI}_i \land \bigwedge_{k \in I} \text{X}([\text{Bl}_i \to k_i] \land [\text{Bl}_i \to k_i])\]
   by 11 and 6, Axioms \text{PerfectInfo} and \text{TimeVert}, Axiom \text{T} for \Box, Axiom \text{K} for \lbrack K \rbrack, and boolean principles;

13. \[\vdash \text{ELEG} \left(\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{m,\text{Ag}}] \text{AX}^m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1}\right)\]
   \[\to \forall i \in \text{Ag}, k \in I, \text{turn}_i \land k_i \land [\lbrack K_i \rbrack] k_i\]
   by 12 and Axiom \text{K} for \lbrack K \rbrack and \text{X};

14. \[\vdash \text{ELEG} \left(\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{m,\text{Ag}}] \text{AX}^m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1}\right)\]
   \[\to \forall i \in \text{Ag}, k \in I, \text{turn}_i \land k_i \land [\lbrack K_i \rbrack] k_i\]
   by 13 and Axiom \text{TimePref};

15. \[\vdash \text{ELEG} \left(\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{m,\text{Ag}}] \text{AX}^m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1}\right)\]
   \[\to [K_j] \text{AX}([K_j] \text{Bl}_i \land \bigwedge_{k \in J} k_i \land \Box([\text{Bl}_i \to k_i] \land [\text{Bl}_i \to k_i]))\]
   by 9 and Lemma A.7;

16. \[\vdash \text{ELEG} \left(\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{m,\text{Ag}}] \text{AX}^m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1}\right)\]
   \[\to [K_j] \text{AX} \left(\bigwedge_{k \in J} k_i \land [K_j] k_i\right)\]
   by 15, Axiom \text{PerfectInfo}, and Axiom \text{K} for \lbrack K \rbrack;

17. \[\vdash \text{ELEG} \left(\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{m,\text{Ag}}] \text{AX}^m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1}\right)\]
   \[\to \forall i \in \text{Ag}, \text{turn}_i \land [\text{Rat}_{i,\text{end}}] \land [K_i] \text{AX} \bigwedge_{k \in I} (k_i \land [K_i] k_i)\]
   by 6 and 16, the definition of \text{Rat}_{i,\text{end}}, Axiom \text{TurnTaking}, and boolean principles;

18. \[\vdash \text{ELEG} \left(\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{m,\text{Ag}}] \text{AX}^m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1}\right)\]
   \[\to \forall i \in \text{Ag}, \text{turn}_i \land [K_i] \text{AX} \bigwedge_{k \in I} (k_i \land [K_i] k_i) \land [K_i] \text{AX} \bigwedge_{h \in I} (h_i \land [K_i] h_i)\]
   by 17, the definition of \text{Rat}_{i,\text{end}};

19. \[\vdash \text{ELEG} \left(\bigwedge_{0 \leq m \leq n+1} [\text{CK}_{m,\text{Ag}}] \text{AX}^m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1}\right)\]
   \[\to \forall i \in \text{Ag}, \text{turn}_i \land [K_i] \text{AX} \bigwedge_{k \in I} (k_i \land [K_i] k_i) \land [K_i] \text{AX} \bigwedge_{h \in I} (h_i \land [K_i] h_i)\]
   by 18, Axiom \text{SinglePref}, and Axiom \text{T} for \lbrack K \rbrack (or Axiom \text{D} if \lbrack K \rbrack is \text{K}D45 modal operator);
20. \( \vdash \text{ELEG} \left( \bigwedge_{0 \leq m \leq n+1} [\text{CK}_m \text{Agt}] \text{AX}_m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \right) \)
   \( \rightarrow \forall i \in \text{Agt} \left( \text{turn}_i \land \forall k \in I \langle K_i \rangle k_i \land \text{AX}(\text{BI}^n \rightarrow \forall h \in I : h \leq k h_i) \right) \)

   by 19 and 9, Lemma A.7, and Axiom \( T \) for \([K_i]\) (or Axiom \( D \) if \([K_i]\) is \( KD45 \) modal operator);

21. \( \vdash \text{ELEG} \left( \bigwedge_{0 \leq m \leq n+1} [\text{CK}_m \text{Agt}] \text{AX}_m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \right) \)
   \( \rightarrow \forall i \in \text{Agt} \left( \text{turn}_i \land \forall k \in I \langle K_i \rangle k_i \land \text{AX}(\text{BI}^n \rightarrow \forall h \in I : h \leq k h_i) \right) \)

   by 20, Axioms \( \text{Perm}_{\square, \text{AX}} \) and \( \text{PerfectInfo} \), and Axioms \( T \) and 5 for \( \square \);

22. \( \vdash \text{ELEG} \left( \bigwedge_{0 \leq m \leq n+1} [\text{CK}_m \text{Agt}] \text{AX}_m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \right) \)
   \( \rightarrow \forall i \in \text{Agt} \left( \text{turn}_i \land \forall k \in I \langle K_i \rangle k_i \land \text{AX}(\text{BI}^n \land \forall h \in I : h \leq k h_i) \right) \)

   by 21 and 6, and Axiom \( K \) for \( \text{AX} \);

23. \( \vdash \text{ELEG} \left( \bigwedge_{0 \leq m \leq n+1} [\text{CK}_m \text{Agt}] \text{AX}_m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \right) \)
   \( \rightarrow \neg \text{end} \land \forall i \in \text{Agt} \left( \text{turn}_i \land \forall k \in I \langle K_i \rangle k_i \land \text{AX}(\forall h \in I : h \leq k h_i \land \text{BI}^n) \right) \)

   by 22 and 14, the definition of \( \text{Depth}^{n+1} \), and boolean principles;

24. \( \vdash \text{ELEG} \left( \bigwedge_{0 \leq m \leq n+1} [\text{CK}_m \text{Agt}] \text{AX}_m \text{AllRat} \land \text{Depth}^{n+1} \land \text{GenPos}^{n+1} \right) \)
   \( \rightarrow \text{BI}^{n+1} \)

   by 23 and the definition of \( \text{BI}^{n+1} \);
Appendix B

Instructions of the Experimental Study

B.1 Preliminary Instructions

We are going to present two games that you will have to play with some unknown participants. One of these games will then be drawn in order to determine your actual winnings.

Each game considers two players. You will be asked to take a decision as Player (1) and as Player (2). At the end of the experiment, we will randomly assign one of these two roles to you.

Your actual payoff will then depend on your decision in the role that will be assigned to you as well as your partner’s decision in the selected game. Therefore, each of your decisions is important. So please take every question seriously by carefully answering them.

Moreover your participation to this experiment relies on the fact that you answered every single question.

If anything is unclear or if you have any question, please do not hesitate to raise your hand so that we can bring you the clarification that you need.

B.2 Instructions of the Baseline game

During this experiment, you will interact with some randomly selected player in a game that is defined as follows.

In the first stage, some initial amount are given to both you and your opponent:

- 20 Euros for Player (1)
- **10 Euros** for Player (2)

No decision needs to be taken by any player during this stage.

In the second stage, every player will then have to choose simultaneously between two distinct moves A and B.

**In the second stage:**

- If every player chooses to play A, 5 Euros will be withdrawn from Player (2)’s initial amount and 15 Euros will be added to Player (1)’s initial amount. Thus Player (1) will get 35 Euros while Player (2) will get 5 Euros.

- If every player chooses to play B, 5 Euros will be withdrawn from Player (1)’s initial amount and 25 Euros will be added to Player (2)’s initial amount. Thus Player (1) will get 15 Euros while Player (2) will get 35 Euros.

- If the players’ choices are different, then both players’ amount will be reset to zero (each will thus get 0 euro).

The following table summarizes the various choices and payoffs from the second stage:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>(1): 35</td>
<td>(1): 0</td>
</tr>
<tr>
<td></td>
<td>(2): 5</td>
<td>(2): 0</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>(1): 0</td>
<td>(1): 15</td>
</tr>
<tr>
<td></td>
<td>(2): 0</td>
<td>(2): 35</td>
</tr>
</tbody>
</table>

This simultaneous decision ends both the second stage and the game. All along the game, both players will remain anonymous to one another. You will receive the corresponding amount if this game is eventually being selected.

These instructions concern the three situations described below.
B.2.1 Questions

In the context of the previous game, you will play with $X$ \(^1\) \textit{(select one answer per question)}

- Please indicate your choice while you are acting as Player (1):
  
  In the second stage, you play:
  \[ \begin{array}{cc}
  \text{O} & A \\
  \text{O} & B \\
  \end{array} \]

- Please indicate your choice while you are acting as Player (2):
  
  In the second stage, you play:
  \[ \begin{array}{cc}
  \text{O} & A \\
  \text{O} & B \\
  \end{array} \]

Note that the three previous questions are independent from one another. Please make sure to answer each of them.

B.3 Instructions of the Entrance game

During this experiment, you will interact with some randomly selected player in a game that is defined as follows.

In the first stage, some initial amount are given to both you and your opponent:

- 20 Euros for Player (1)
- 10 Euros for Player (2)

Then, the two following options become available to Player (1):

- The “Out” option implies that every player keeps their initial amount and the game ends.
- The alternative option (“In”) implies entering a second stage where each player will have to take another decision. In the latter case, both players will then have to choose simultaneously between two distinct moves A and B.

\(^1\)Depending on the matching process, $X$ may stand for “a university student”, “a club member”, or “a teammate” (See Section 5.2 for details about the matching process).
In the second stage:

- If every player chooses to play A, 5 Euros will be withdrawn from Player (2)'s initial amount and 15 Euros will be added to Player (1)'s initial amount. Thus Player (1) will get 35 Euros while Player (2) will get 5 Euros.

- If every player chooses to play B, 5 Euros will be withdrawn from Player (1)'s initial amount and 25 Euros will be added to Player (2)'s initial amount. Thus Player (1) will get 15 Euros while Player (2) will get 35 Euros.

- If the players' choices are different, then both players' amount will be reset to zero (each will thus get 0 euro).

The following table summarizes the various choices and payoffs from the second stage:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1): 35</td>
<td>(1): 0</td>
</tr>
<tr>
<td>A</td>
<td>(2): 5</td>
<td>(2): 0</td>
</tr>
<tr>
<td>B</td>
<td>(1): 0</td>
<td>(1): 15</td>
</tr>
<tr>
<td></td>
<td>(2): 0</td>
<td>(2): 35</td>
</tr>
</tbody>
</table>

This simultaneous decision ends both the second stage and the game. All along the game, both players will remain anonymous to one another. You will receive the corresponding amount if this game is eventually being selected.

These instructions concern the three situations described below.
B.3.1 Questions

In the context of the previous game, you will play with $X$ \(^1\) \textit{(select one answer per question)}

- Please indicate your choice while you are acting as \textbf{Player (1)}:

  In the first stage, you play:
  
  - $\bigcirc$ In
  - $\bigcirc$ Out

  In the second stage (assume that you played “In” first), you play:
  
  - $\bigcirc$ A
  - $\bigcirc$ B

- Please indicate your choice while you are acting as \textbf{Player (2)}:

  In the second stage (assume that your opponent played “In” first), you play:
  
  - $\bigcirc$ A
  - $\bigcirc$ B

Note that the three previous questions are independent from one another. Please make sure to answer each of them.

B.4 Belief questions

B.4.1 In the Baseline game

In a previous study concerning this game, 20 students from Toulouse 1 university capitole were randomly selected. Among these, 10 subjects played as \textit{player (1)} and 10 subjects played as \textit{player (2)}. It was common knowledge among all these subjects that they were interacting between student from the same university.

In this context, we ask you to indicate what you believe these people did actually play. Each of the two following answers (one for each player) that matches the actual behavior will earn you 5 euros (after each proposition, please indicate a number ranging from 0 to 10).

- \textit{(5 euros)} Out of the 10 players (1), indicate how many did choose A: ______ and B: ______

\(^1\)Depending on the matching process, $X$ may stand for “a university student”, “a club member”, or “a teammate” (See Section 5.2 for details about the matching process).
• (5 euros) Out of the 10 players (2), indicate how many did choose A: ______ and B: ______

B.4.2 In the Entrance game

In a previous study concerning this game, 20 students from Toulouse 1 university capitole were randomly selected. Among these, 10 subjects played as player (1) and 10 subjects played as player (2). It was common knowledge among all these subjects that they were interacting between student from the same university.

In this context, we ask you to indicate what you believe these people did actually play. Each of the two following answers (one for each player) that matches the actual behavior will earn you 5 euros (after each proposition, please indicate a number ranging from 0 to 10).

• (5 euros) Out of the 10 players (1), indicate how many did choose “In”: ______
   Out of these subjects, how many did then choose A: ______ and B: ______

• (5 euros) Out of the 10 players (2), indicate how many did choose A: ______ and B: ______
B.5 Elicited beliefs

Figure B.1: Beliefs about Player (1)’s actual choice in the Baseline game

Figure B.2: Beliefs about Player (2)’s actual choice in the Baseline game
(a) Belief about playing In/Out

(b) Belief about playing \((In, A)/(In, B)\)

Figure B.3: Beliefs about Player (1)’s actual choice in the Entrance game
Figure B.4: Beliefs about Player (2)'s actual choice in the Entrance game
Appendix C

Social ties and team reasoning

C.1 Proof of Theorem 6.2

We here demonstrate that, given a strategic game with group utility \( G = (\{i, j\}, \{S_i, S_j\}, \{U_i, U_j, U_{\{i,j\}}\}) \) where \( \text{argmax}_{s \in S} U_{\{i,j\}}(s) \) is a singleton, the predictions made by both game structures \( \text{BST} = (G, \{k_i, k_j\}) \) and \( \text{BUTI} = (G, \{\Omega_i, \Omega_j\}) \) are equivalent whenever \( k_i(\{i, j\}) = k_j(\{i, j\}) = \Omega_i(\{i, j\}) = \Omega_j(\{i, j\}) \).

For this purpose, one can distinguish between two different cases.

If \( k_i(\{i, j\}) = k_j(\{i, j\}) = \Omega_i(\{i, j\}) = \Omega_j(\{i, j\}) = 0 \), then both agents \( i \) and \( j \) reason in I-mode, i.e., \( k_i(\{i\}) = k_j(\{j\}) = \Omega_i(\{i\}) = \Omega_j(\{j\}) = 1 \). In this case, for any protocol \( \alpha \in \Delta \), the value of \( \alpha^{\{i,j\}} \) has therefore no influence on finding an UTI equilibrium in \( \text{BUTI} \). Moreover, it is easy to show, through Definition 6.7, that the game induced by \( \text{BST} \) corresponds to the original game \( G \). Following Definition 6.3, given \( s \in S \) and \( \alpha \in \Delta \) such that \( s = (\alpha^{\{i\}}, \alpha^{\{j\}}) \), we then have that \( s \) is a Nash equilibrium in \( G \) if and only if \( \alpha \) is an UTI equilibrium in \( \text{BUTI} \).

Let us similarly consider the case where \( k_i(\{i, j\}) = k_j(\{i, j\}) = \Omega_i(\{i, j\}) = \Omega_j(\{i, j\}) = 1 \), which implies that both agents \( i \) and \( j \) reason in we-mode.

From Definitions 6.9 and 6.6, it follows that, in \( \text{BST} \), for every \( s \in S \), we have:

\[
(1) \quad U_i^{ST}(s) = \max_{s_j' \in S_j} U_{\{i,j\}}(s_i, s_j')
\]

and

\[
(2) \quad U_j^{ST}(s) = \max_{s_i' \in S_i} U_{\{i,j\}}(s_j, s_i')
\]

226
Then (1) and (2) allow to state that, given \( s \in S \) is a Nash equilibrium solution in the strategic game induced by BST, then we have:

\[
U^S_i(s) = U^S_j(s) = U_{\{i,j\}}(s) = \max_{s' \in S} U_{\{i,j\}}(s')
\]

Similarly considering the BUTI structure, it follows from Definition 6.8 that for the team \{i, j\} and every protocol \( \alpha \in \Delta \), the expected value can be simplified as follows:

\[
EV_{\{i,j\}}(\alpha) = U_{\{i,j\}}(\alpha(i, \{i,j\}), \alpha(j, \{i,j\})) = U_{\{i,j\}}(\alpha^{(i,j)})
\]

In fact, for any protocol \( \alpha \in \Delta \), the values of \( \alpha^{(i)} \) and \( \alpha^{(j)} \) are then no influence on determining the group’s expected value.

Moreover, according to (4) and Definition 6.3, \( \alpha \in \Delta \) is an UTI equilibrium if and only if:

\[
EV_{\{i,j\}}(\alpha) = \max_{s' \in S} U_{\{i,j\}}(s')
\]

It then follows from (3) and (5) that, for some \( s \in S \) and \( \alpha \in \Delta \) such that \( s = \alpha^{(i,j)} \), if argmax\( s' \in S U_{\{i,j\}}(s') \) is a singleton, then \( s \) is a Nash equilibrium in the game induced by BST if and only if \( \alpha \) is an UTI equilibrium in BUTI.

### C.2 Gradual group identification

We consider a two-player strategic game with group utility \( G = \langle \{i, j\}, \{S_i, S_j\}, \{U_i, U_j, U_{\{i,j\}}\} \rangle \) where \( S_i = \{A, B, C\} \) and \( S_j = \{D\} \), and the individual payoff matrix is represented in Table 6.4 from Section 6.3.3.1 for both players from \{i, j\}. Moreover, we consider as the collective payoff function either the classical utilitarian principle (i.e., for any \( s \in S \), \( U_{\{i,j\}}(s) = U_i(s) + U_j(s) \)) or the Rawlsian criterion of fairness (i.e., for any \( s \in S \), \( U_{\{i,j\}}(s) = \min(U_i(s), U_j(s)) \)).

The obvious main characteristics of the game depicted in Table 6.4 is that the outcome is uniquely determined by a single player (i.e., player i).

We then define a corresponding social ties game \( ST = \langle G, \{k_i, k_j\} \rangle \) such that \( k_i(\{i,j\}) = k_j(\{i,j\}) = 0.5 \). One should then note that, in order to satisfy Constraint C4 from Definition 6.5, we must have that \( k_i(\{i\}) = k_j(\{j\}) = 0.5 \). The resulting strategic game induced by \( ST \) when considering the classical utilitarian principle as the collective payoff function is depicted in Figure C.1(a) (in Figure C.1, player i corresponds to the row player while player j corresponds to the column player). Similarly, the strategic game induced by \( ST \) when considering Rawls’ principle of fairness as the collective payoff function is depicted in Figure C.1(b).
In this case, it is straightforward to show that in both games from Figure C.1, the best strategy for player $i$ is to play $C$.

On the other hand, let us now consider a corresponding structure $UTI = \langle G, \{\Omega_i,\Omega_j}\rangle$ such that $\omega$ defines the probability that player $i$ reasons in _we-mode_, that is $\omega = \Omega_i(\{i,j\})$ and $\omega = 1 - \Omega_i(\{i\}) = 1$.

Table C.1 illustrates the expected values for every possible protocol in the game from Table 6.4. Note that two different interpretations of the group’s expected value is provided depending on the type of collective payoff function: $EV^u_{\{i,j\}}$ is calculated based on the classical utilitarian principle whereas $EV^m_{\{i,j\}}$ is calculated based on the Rawlsian principle of fairness (i.e., _maximin_).

<table>
<thead>
<tr>
<th>Protocols</th>
<th>Expected values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_n$, $\alpha_n(a, {a})$, $\alpha_n(a, {a, b})$</td>
<td>$EV_u(\alpha_n)$, $EV^u_{{a,b}}(\alpha_n)$, $EV^m_{{a,b}}(\alpha_n)$</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>$A$</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>$B$</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>$C$</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>$A$</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>$A$</td>
</tr>
<tr>
<td>$\alpha_6$</td>
<td>$B$</td>
</tr>
<tr>
<td>$\alpha_7$</td>
<td>$B$</td>
</tr>
<tr>
<td>$\alpha_8$</td>
<td>$C$</td>
</tr>
<tr>
<td>$\alpha_9$</td>
<td>$C$</td>
</tr>
</tbody>
</table>

Table C.1: All protocols in the dictator game (see Table 6.4)

It is then straightforward to show from Table C.1 and Definition 6.3 that, for

\[\text{Note that the probability function } \Omega_j \text{ is irrelevant here as player } j \text{ cannot influence the outcome of the game.}\]
any value of $\omega$ such that $0 < \omega < 1$, the structure UTI yields the unique UTI equilibrium $\alpha_4$. Indeed, for every $0 < \omega < 1$, we have:

$$
EV_{i,i}(\alpha_4) > EV_{i,i}(\alpha_2) > EV_{i,i}(\alpha_5) > EV_{i,i}(\alpha_7) \\
EV_{i,i}(\alpha_5) > EV_{i,i}(\alpha_3) > EV_{i,i}(\alpha_1) > EV_{i,i}(\alpha_8) \\
EV_{i,i}(\alpha_1) > EV_{i,i}(\alpha_6) > EV_{i,i}(\alpha_4) > EV_{i,i}(\alpha_9) \\
EV_{i,i}(\alpha_4) > EV_{i,i}(\alpha_6) > EV_{i,i}(\alpha_5) > EV_{i,i}(\alpha_7) \\
EV_{i,i}(\alpha_5) > EV_{i,i}(\alpha_3) > EV_{i,i}(\alpha_1) > EV_{i,i}(\alpha_8)
$$

Moreover, player $i$ will play $A$ whenever $\omega = 0$ whereas player $i$ will play $B$ whenever $\omega = 1$. As a result, there exist no UTI structure built on the game $G$ that has an UTI equilibrium specifying player $i$ to play $C$, as predicted by the above social ties game $ST$.

### C.3 Proof of Theorem 6.3

We here demonstrate that, given a strategic game with group utility $G = \langle Agt, \{S_i|i \in Agt\}, \{U_J|J \in 2^{Agt^*}\} \rangle$ that satisfies Constraint C3 from Definition 6.10, both game structures $BST = \langle G, \{k_i|i \in Agt\} \rangle$ and $BUTI = \langle G, \{\Omega_i|i \in Agt\} \rangle$ can be transformed respectively into simpler equivalent strategic games $G^{bst}$ and $G^{buti}$.

Let us first consider the $BST$ game.

From Definitions 6.9 and 6.6, it follows that, for every $i \in Agt$, $s \in S$, and for some $J \subseteq Agt \setminus \{i\}$:

$$
U_{i,J}(s) = \max_{s^{'}_J \in S_J} U_{J \cup \{i\}}(s^{'}_J, s^{'}_J)
$$

From (1), Definition 6.5 and Constraint C2 allow to state that, for every agent $j \in J$ such that $j \neq i$:

$$
U_{j,J}(s) = \max_{s^{'}_K \in S_K} U_{K \cup \{j\}}(s^{'}_K, s^{'}_K)
$$

where $K = J \cup \{i\} \setminus \{j\}$

Then (1) and (2) allow to state that, given $s \in S$ is a Nash equilibrium solution in the strategic game induced by $BST$, then for every coalition $J \in C$ and every agent $i \in J$, we have:

$$
U_{i,J}(s) = U_{J}(s) = \max_{s^{'}_J \in S_J} U_{J}(s^{'}_J, s^{'}_J)
$$

Moreover, let $C$ denote the set of actual coalitions in the $BST$ game:

$$
C = \{J \in 2^{Agt^*}|\forall i \in J, k_i(J) = 1\}
$$
Note that \( \bigcup_{J \in C} J = \text{Agt} \) and \( \bigcap_{J \in C} = \emptyset \).

As a result from (3) and (4), the current game \( BST = \langle \text{Agt}, \{ S_i | i \in \text{Agt} \}, \{ U_J | J \in 2^{\text{Agt}^*} \}, \{ k_i | i \in \text{Agt} \} \rangle \) can be transformed into a simplified strategic game \( G^{bst} = \langle \text{Agt}', \{ S'_J | J \in \text{Agt}' \}, \{ U'_J | J \in \text{Agt}' \} \rangle \) where each group \( J \in C \) acts as a single agent, that is:

- \( \text{Agt}' = C \);
- for every \( J \in C \), \( S'_J = S_J \);
- for every \( J \in C \) and every \( s \in \prod_{K \in C} S'_K, U'_J(s) = U_J(s) \).

Note that through this game transformation, \( 1 \leq |\text{Agt}'| \leq |\text{Agt}| \). In the particular case where every agent in \( \text{Agt} \) is individualistic (i.e., \( k_i(\{i\}) = 1 \) for every \( i \in \text{Agt} \)), then \( |\text{Agt}'| = |\text{Agt}| \).

It is then straightforward to show that, as Constraint C3 from Definition 6.10 ensures that every group \( J \in C \) has no conflicting goals in \( G \) (and in \( G^{bst} \)), it implies that \( s \in S \) is a Nash equilibrium in \( G^{bst} \) if and only if \( s \) is a Nash equilibrium in the game induced by \( BST \).

Let us now similarly consider the BUTI structure.

Definition 6.8 means that every agent can only identify with a unique group. It follows from this definition that there exists a unique group identification state \( g \in \text{Groups} \) such that:

\[
\bigcup_{i \in \text{Agt}} g_i = \text{Agt} \quad \text{and} \quad \bigcap_{i \in \text{Agt}} g_i = \emptyset
\]

Given the unique group identification state \( g \) that satisfies (5), one can therefore define the set of active teams \( A \) as follows:

\[
A = \{ J \in 2^{\text{Agt}^*} | \forall i \in J, g_i = J \}
\]

It follows from (6) and Definition 6.8 that for every team \( J \in 2^{\text{Agt}^*} \) and every protocol \( \alpha \in \Delta \), the expected value can be simplified as follows:

\[
EV_J(\alpha) = U_J(\alpha(1, g_1), \ldots, \alpha(n, g_n))
\]

One can note from (7) and Definition 6.3 that, if protocol \( \alpha \) is an \( \text{UTI} \) equilibrium, then, for every \( J \in A \), we have:

\[
EV_J(\alpha) = \max_{s_J' \in S_J} U_J(s_J', s_{-J})
\]

where \( s_{-J} = \prod_{J' \in A : J' \neq J} \alpha^{J'} \) is the combination of actions of all agents \( i \in \text{Agt} \setminus J \) specified by protocol \( \alpha \) when identifying with any group from \( A \). Note that, for
every coalition $J \in 2^{Agt}$ such that $J \notin A$ (i.e., for every team nobody identifies with), and for every $\alpha \in \Delta$, $EV_J(\alpha)$ remains constant. This implies that every such group $J$ has simply no influence on the computation of an UTI equilibrium.

As a result from (5), (6), (7) and (8), a structure $BUTI = \langle Agt, \{S_i|i \in Agt\}, \{U_J|J \in 2^{Agt}\}, \{\Omega_i|i \in Agt\} \rangle$ can be transformed into a simplified strategic game $G^{buti} = \langle Agt'', \{S''_i|J \in Agt''\}, \{U''_J|J \in Agt''\} \rangle$ where every active team $J \in A$ acts as a single agent, that is:

- $Agt'' = A$;
- for every $J \in A$, $S''_J = S_J$;
- for every $J \in A$ and every $s \in \prod_{J \in A} S''_J$, $U''_J(s) = U_J(s)$.

Note that through this game transformation, $1 \leq |Agt''| \leq |Agt|$. In the particular case where every agent is individualistic (i.e., $g_i = \{i\}$ for every $i \in Agt$), then $|Agt''| = |Agt|$.

Given $s \in S$, let $\alpha \in \Delta$ be a protocol such that, for every $J \in 2^{Agt}$, $\alpha^J = s_J$. It is then straightforward to demonstrate from (8) and Definition 6.3 that $s$ is a Nash equilibrium in $G^{buti}$ if and only if $\alpha$ is an UTI equilibrium in $BUTI$. Note however that this equivalence does here not rely on Constraint C3 from Definition 6.10.

Finally, given that for every $i \in Agt$ and every $J \in Group(i)$, $k_i(J) = \Omega_i(J)$, it follows from (4) and (6) that $A = C$ and consequently $G^{buti} = G^{bst}$. As a result, given $s \in S$ and $\alpha \in \Delta$ such that, for every $J \in 2^{Agt}$, $\alpha^J = s_J$, it implies that $s$ is a Nash equilibrium in the game induced by $BST$ if and only if $\alpha$ is an UTI equilibrium in $BUTI$.

### C.4 Utility transformation vs. agency transformation

We consider a three-player strategic game with group utility $G = \langle Agt, \{S_i|i \in Agt\}, \{U_J|J \in 2^{Agt}\} \rangle$ where $Agt = \{1, 2, 3\}$, $S_i = \{A, B\}$ for every $i \in Agt$, and the individual payoff matrix is represented in Table 6.5 from Section 6.3.3.3 for all players from $Agt$. Moreover, we take as the collective payoff function the Rawlsian criterion of fairness (i.e., for every $J \in 2^{Agt}$ and every $s \in S$, $U_J(s) = \min_{i \in J} U_i(s)$).

The main characteristics of this game is that $G$ does not satisfy Constraint C3 from Definition 6.10. In fact, if player 3 selects $A$, then the team $\{1, 2\}$ made of the two other players appears to have conflicting goals, that is, both $(A, A, A)$ and $(B, B, A)$ are equally the best options for the group $\{1, 2\}$. 

231
Let us now define a corresponding structure $\text{BUTI} = \langle G, \{\Omega_i | i \in \text{Agt}\} \rangle$ such that $\Omega_1(\{1,2\}) = \Omega_2(\{1,2\}) = 1$, and $\Omega_1(\{1,3\}) = \Omega_2(\{1,3\}) = \Omega_2(\{2,3\}) = \Omega_3(\{2,3\}) = \Omega_3(\{1,2,3\}) = 0$.

In this case, it is straightforward to show that every protocol $\alpha \in \Delta$ that is an $\text{UTI}$ equilibrium is such that $\alpha(\{1,2\}) = (A,A)$.

Similarly, we define a binary social ties game $\text{BST} = \langle G, \{k_i | i \in \text{Agt}\} \rangle$ and determine the various predictions such a game can make depending on the valuation of the functions $k_i$. The corresponding sets of Nash equilibria in the game induced by $\text{BST}$ can be found in Table C.2.

<table>
<thead>
<tr>
<th>$k_i({1,2})$</th>
<th>$k_i({1,3})$</th>
<th>$k_i({2,3})$</th>
<th>$k_i({1,2,3})$</th>
<th>Predicted outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$(A,A,A)$</td>
</tr>
<tr>
<td></td>
<td>$(A,A,B)$</td>
<td></td>
<td></td>
<td>$(B,B,B)$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$(A,A,A)$</td>
</tr>
<tr>
<td></td>
<td>$(A,A,B)$</td>
<td></td>
<td></td>
<td>$(A,B,A)$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$(B,B,B)$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$(A,A,A)$</td>
</tr>
<tr>
<td></td>
<td>$(A,A,B)$</td>
<td></td>
<td></td>
<td>$(B,B,B)$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>$(B,A,A)$</td>
</tr>
</tbody>
</table>

Table C.2: A three-player coordination game

One can therefore observe from Table C.2 that there exists no $\text{BST}$ structure such that $(A,A,B)$ is the unique Nash equilibrium in the game induced by $\text{BST}$.

### C.5 Forming sub-coalitions in the Three Musketeers game

We here analyse in more details the Three Musketeers game presented in Section 6.3.3 (see the payoff matrix in Table 6.6), depending on the settings of the players’ social ties. To do so, let us define a strategic game with group utility $G = \langle \text{Agt}, \{S_i | i \in \text{Agt}\}, \{U_J | J \in 2^{\text{Agt}}\} \rangle$ such that $\text{Agt} = \{1,2,3\}$ (we assume that players 1, 2, and 3 respectively stand for Athos, Porthos, and Aramis), and functions $S_i$ and $U_J$ are defined according to Table 6.6 and Rawls’ criterion of fairness (i.e., $\text{maximin}$).

\[^1]\text{Note that } \alpha^{(3)} \text{ may then play either } A \text{ or } B, \text{ and every other group } J \in 2^{\text{Agt}} \text{ such that } J \neq \{1,2\} \text{ and } J \neq \{3\} \text{ is irrelevant as } J \text{ is an inactive team.}
Moreover, in the following analysis, we strictly focus on the most relevant situation where two sub-coalitions (whose intersection is non-empty) may be formed within the game. More specifically, we only consider situations where identifying with the largest coalition is negligible compared to identifying with any sub-coalition.

C.5.1 Prediction in a social ties game

Let us define a social ties game $ST = \langle G, \{k_i| i \in Agt}\rangle$ where the function $k$ is defined as follows:

- $k_1(\{1, 2, 3\}) = k_2(\{1, 2, 3\}) = k_3(\{1, 2, 3\}) = k_2(\{2, 3\}) = k_3(\{2, 3\}) = k_1(\{1\}) = 0$;
- $k_1(\{1, 2\}) = k_1(\{1, 3\}) = k_2(\{2\}) = k_3(\{3\}) = 0.5$.

The most notable characteristics of this scenario is that it creates two coalitions whose intersection is non-empty: the team $\{1, 2\}$ is made of Athos and Porthos while the team $\{1, 3\}$ is made of Athos and Aramis. More specifically, Porthos and Aramis are torn between satisfying their respective coalition (resp. groups $\{1, 2\}$ and $\{1, 3\}$) and maximizing their own self-interest (resp. groups $\{2\}$ and $\{3\}$). On the other hand, Athos is torn between satisfying either coalitions he identifies with (in this case, Athos is not driven by any self-interest, i.e., $k_1(\{1\}) = 0$). Furthermore, one should note that, as a consequence of the absence of any tie between Porthos and Aramis, this scenario induces no identification with the large group $\{1, 2, 3\}$ (i.e., $k_i(\{1, 2, 3\}) = 0$ for every $i \in \{1, 2, 3\}$).

The corresponding transformed game taking into account each player’s Social Ties utility based on the previous parameters and Definition 6.6 is found in Table C.3.

It is therefore straightforward to show through Table C.3 that the unique Nash equilibrium in the transformed game is $(C, C, C)$. In other words, even though the players are not unified as a unique coalition, our theory predicts that they will manage to coordinate on the same outcome.

C.5.2 Prediction in an unreliable team interaction structure

Let us now perform a similar analysis of this game through Bacharach’s theory of team reasoning. Our aim here is to demonstrate that interpreting social ties within some unreliable team interaction structure cannot yield any stable solution where all individuals select $C$ in the absence of a unique strong coalition (i.e., where the probability of identifying with the largest group is negligible).
In order to allow for the comparison with the previous ST game, we define a corresponding structure \( UTI = \langle G, \{ \Omega_i | i \in \text{Agt} \} \rangle \) where we assume that an agent identifies with a group if and only if other members of that group also identify with it. Thus, let \( \omega_{12}, \omega_{13}, \) and \( \omega_{23} \) denote respectively the probability of identifying with the coalition \( \{1, 2\} \) (i.e., \( \Omega_1(\{1, 2\}) = \Omega_2(\{1, 2\}) = \omega_{12} \)), the probability of identifying with the coalition \( \{1, 3\} \) (i.e., \( \Omega_1(\{1, 3\}) = \Omega_3(\{1, 3\}) = \omega_{13} \)), and the probability of identifying with the coalition \( \{2, 3\} \) (i.e., \( \Omega_2(\{2, 3\}) = \Omega_3(\{2, 3\}) = \omega_{23} \)). Similarly, the probability of identifying with the largest group is defined by \( \omega_{123} \) (i.e., \( \Omega_1(\{1, 2, 3\}) = \Omega_2(\{1, 2, 3\}) = \Omega_3(\{1, 2, 3\}) = \omega_{123} \)). In order to satisfy Definition 6.3, we recall that the following constraints must hold:

\[
\begin{align*}
(1) & \quad \omega_{12} + \omega_{13} + \omega_{123} \leq 1 \\
(2) & \quad \omega_{12} + \omega_{23} + \omega_{123} \leq 1 \\
(3) & \quad \omega_{13} + \omega_{23} + \omega_{123} \leq 1
\end{align*}
\]

Moreover, one should note that such a representation implies that \( \Omega_1(\{1\}) = 1-\omega_{12}-\omega_{13}-\omega_{123}, \Omega_2(\{2\}) = 1-\omega_{12}-\omega_{23}-\omega_{123}, \) and \( \Omega_3(\{3\}) = 1-\omega_{13}-\omega_{23}-\omega_{123} \).

In this context, let us define nine protocols \( \alpha_1 \ldots \alpha_9 \) in \( UTI \) that are relevant to our analysis. Detailed specifications of these protocols can be found in Table C.4. Our aim is to demonstrate that protocol \( \alpha_1 \) is necessarily an \( UTI \) equilibrium according to Definition 6.3 if \( \omega_{12} > \omega_{123} \) and \( \omega_{13} > \omega_{123} \) (i.e., the probability of identifying with the largest group is negligible compared to that of identifying with any sub-coalition). In order to do so, one needs to evaluate whether one of the individuals or one of the teams is better off deviating. Note that, according to

<table>
<thead>
<tr>
<th>Actions</th>
<th>Utilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athos</td>
<td>Porthos</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Table C.3: The transformed game with updated utilities
### Table C.4: Various protocols in the Three Musketeers game

<table>
<thead>
<tr>
<th>Protocols</th>
<th>Expected Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha_1)</td>
<td>(EV_{(1)}(\alpha_1) = 5)</td>
</tr>
<tr>
<td></td>
<td>(EV_{(2)}(\alpha_1) = 5(1 - \omega_{13} - \omega_{123}) + 4\omega_{123})</td>
</tr>
<tr>
<td></td>
<td>(EV_{(3)}(\alpha_1) = 5\omega_{13} + 4\omega_{123})</td>
</tr>
<tr>
<td></td>
<td>(EV_{(1,2)}(\alpha_1) = 5(1 - \omega_{13} - \omega_{123}) + 4\omega_{123})</td>
</tr>
<tr>
<td></td>
<td>(EV_{(1,3)}(\alpha_1) = 5\omega_{13} + \omega_{123})</td>
</tr>
<tr>
<td></td>
<td>(EV_{(2,3)}(\alpha_1) = 4\omega_{123})</td>
</tr>
<tr>
<td></td>
<td>(EV_{(1,2,3)}(\alpha_1) = 4\omega_{123})</td>
</tr>
<tr>
<td>(\alpha_2)</td>
<td>(EV_{(2)}(\alpha_2) = 4(\omega_{12} + \omega_{123})\omega_{123})</td>
</tr>
<tr>
<td></td>
<td>+5(\omega_{12} + \omega_{123})(1 - \omega_{13} - \omega_{123})</td>
</tr>
<tr>
<td></td>
<td>+6\omega_{123}(1 - \omega_{12} - \omega_{123})</td>
</tr>
<tr>
<td>(\alpha_3)</td>
<td>(EV_{(3)}(\alpha_3) = 6\omega_{123}(1 - \omega_{13} - \omega_{123}))</td>
</tr>
<tr>
<td></td>
<td>+ (\omega_{13} + \omega_{123})(4\omega_{123} + 5\omega_{13})</td>
</tr>
<tr>
<td>(\alpha_4)</td>
<td>(EV_{(1,2)}(\alpha_4) = (\omega_{12} + \omega_{123})4)</td>
</tr>
<tr>
<td></td>
<td>+5(1 - \omega_{12})(1 - \omega_{12} - \omega_{13} - \omega_{123})</td>
</tr>
<tr>
<td>(\alpha_5)</td>
<td>(EV_{(1,2)}(\alpha_5) = 4(\omega_{12} + \omega_{123}))</td>
</tr>
<tr>
<td></td>
<td>+5(1 - \omega_{12} - \omega_{13} - \omega_{123})</td>
</tr>
<tr>
<td>(\alpha_6)</td>
<td>(EV_{(1,3)}(\alpha_6) = 4(\omega_{13} + \omega_{123}))</td>
</tr>
<tr>
<td>(\alpha_7)</td>
<td>(EV_{(1,3)}(\alpha_7) = 4(\omega_{13} + \omega_{123}))</td>
</tr>
<tr>
<td>(\alpha_8)</td>
<td>(EV_{(2,3)}(\alpha_8) = 3\omega_{123}\omega_{23}^2 + 4\omega_{123}(1 - \omega_{23})^2)</td>
</tr>
<tr>
<td>(\alpha_9)</td>
<td>(EV_{(1,2,3)}(\alpha_9) = 3\omega_{123}^2 + 4\omega_{123}(1 - \omega_{123})^2)</td>
</tr>
</tbody>
</table>

### Table C.5: Expected values of protocols \(\alpha_1 \ldots \alpha_9\)
$\alpha_1$, Porthos will always choose $A$, Aramis will always choose $B$, and Athos will select $C$ if identifying with the largest group, $B$ if identifying with the sub-coalition including only Aramis, and $A$ otherwise. In the following analysis, we will then refer to Table C.5, which provides a set of relevant expected values of each protocol from Table C.4.

Let us first focus on player 1 (i.e., Athos). It is straightforward to show that the expected value of $\alpha$ for Athos is maximal (i.e., according to Table C.5, $EV_{\{1\}}(\alpha_1) = 5$), independently of the probabilities $\omega_{12}$, $\omega_{13}$, $\omega_{23}$, and $\omega_{123}$ (Athos is certain about both Porthos and Aramis’s actions). This implies that Athos has no personal interest for deviating from $\alpha_1$.

Now, concerning player 2 (i.e., Porthos), protocol $\alpha_1$ indicates that he is uncertain about Athos’s action. In order to evaluate whether Porthos should deviate from $\alpha_1$ (i.e., by playing $C$ instead of $A$), let us consider protocol $\alpha_2$. According to Table C.5, it is straightforward to show that $EV_{\{2\}}(\alpha_2) > EV_{\{2\}}(\alpha_1)$ if and only if the following condition holds:

$$ \omega_{123} > \frac{5}{7} \cdot (1 - \omega_{13}) $$

Performing a similar analysis for player 3 (i.e., Aramis) leads us to consider the alternative protocol $\alpha_3$, as a means to verify whether Aramis can personally benefit by deviating from $\alpha_1$ (i.e., by playing $C$ instead of $B$). In this case, According to Table C.5, one can show that $EV_{\{3\}}(\alpha_3) > EV_{\{3\}}(\alpha_1)$ if and only if the following condition holds:

$$ \omega_{123} > \frac{5}{2} \cdot \omega_{13} $$

So far, we have stated under which conditions each of the three players can increase their individual expected value by deviating unilaterally from protocol $\alpha_1$. However, it remains to be shown under which conditions they can benefit by deviating from $\alpha_1$ while they team-reason.

Let us then consider the team $\{1, 2\}$ made of Athos and Porthos. Note from Table C.5 that the team’s expected value of the protocol $\alpha_1$ is the same as that of Porthos alone, that is, $EV_{\{1, 2\}}(\alpha_1) = EV_{\{2\}}(\alpha_1)$. As before, in order to determine whether the team, as a whole, could be better off deviating from $\alpha_1$, let us define any possible alternative protocol. One can then note that the only relevant protocols are $\alpha_4$ and $\alpha_5$ from Table C.4.

In fact, it is straightforward to show from Table 6.6 that any other protocol $\beta$ such that $\beta_{\{1, 2\}} \neq (A, A)$ and $\beta_{\{1, 2\}} \neq (C, \cdot)$ would always yield a lower expected value for the team than $\alpha_1$ (a team is indeed always better off coordinating).

236
In this case, it can then be shown from Table C.5 that $\alpha_4$ outperforms $\alpha_1$ for the team $\{1, 2\}$ if only if the following holds:

\begin{equation}
\omega_{12} + \omega_{13} + \omega_{123} > \frac{6}{5}
\end{equation}

As (6) is in obvious contradiction with (1), this therefore implies that the team $\{1, 2\}$ can never benefit by playing $(C, C)$ instead of $(A, A)$ (i.e., $EV_{\{1,2\}}(\alpha_1) \geq EV_{\{1,2\}}(\alpha_4)$ for any values of probabilities $\omega_{12}, \omega_{13}, \omega_{23}$, and $\omega_{123}$). Moreover, it is also straightforward to show from Table 6.6 that the team $\{1, 2\}$ can also not benefit by playing $(C, A)$ instead of $(A, A)$ (i.e., $EV_{\{1,2\}}(\alpha_1) \geq EV_{\{1,2\}}(\alpha_5)$ for any values of probabilities $\omega_{12}, \omega_{13}, \omega_{23}$, and $\omega_{123}$). As a result, one can states that the protocol $\alpha_1$ is optimal for the team $\{1, 2\}$.

Let us perform the same analysis to the other team $\{1, 3\}$ made of Athos and Aramis. Note, from Table C.5, that the team’s expected value for $\alpha_1$ happens to be the same as that of Aramis alone, that is, $EV_{\{1,3\}}(\alpha) = EV_{\{3\}}(\alpha)$. Furthermore, as in the previous case, we consider the only relevant alternative protocols $\alpha_6$ and $\alpha_7$ from Table C.4\textsuperscript{1}.

In this case, it can then be shown from Table C.5 that $\alpha_6$ outperforms $\alpha_1$ for the team $\{1, 3\}$ if only if the following holds:

\begin{equation}
\omega_{13} + \omega_{123} > 1
\end{equation}

As (7) is in obvious contradiction with (1) and (2), this therefore implies that the team $\{1, 3\}$ can never benefit by playing $(C, C)$ instead of $(B, B)$ (i.e., $EV_{\{1,3\}}(\alpha_1) \geq EV_{\{1,3\}}(\alpha_6)$ for any values of probabilities $\omega_{12}, \omega_{13}, \omega_{23}$, and $\omega_{123}$).

Furthermore, it is straightforward to show that protocol $\alpha_7$ can never strictly outperform $\alpha_1$ for the team $\{1, 3\}$, that is, $EV_{\{1,3\}}(\alpha_1) \geq EV_{\{1,3\}}(\alpha_7)$ for any values of probabilities $\omega_{12}, \omega_{13}, \omega_{23}$, and $\omega_{123}$). In other words, one can states that the protocol $\alpha_1$ is also optimal for the team $\{1, 3\}$.

Applying the same analysis to the team $\{2, 3\}$ made of Porthos and Aramis leads us, as before, to consider the only relevant alternative protocol $\alpha_8$ from Table C.4\textsuperscript{2}. In this case, it is straightforward to show that protocol $\alpha_8$ can never strictly outperform $\alpha_1$ for the team $\{2, 3\}$, that is, $EV_{\{2,3\}}(\alpha_1) \geq EV_{\{2,3\}}(\alpha_8)$ for any values

\textsuperscript{1}As in the previous case, it is straightforward to show from Table 6.6 that any other protocol $\beta'$ such that $\beta'_{\{1,3\}} \neq (B, B)$ and $\beta'_{\{1,3\}} \neq (C, C)$ would always yield a lower expected value for the team than $\alpha_1$.

\textsuperscript{2}Here again, it is straightforward to show from Table 6.6 that any other protocol $\beta''$ such that $\beta''_{\{2,3\}} \neq (A, B)$ and $\beta''_{\{2,3\}} \neq (C, C)$ would always yield a lower expected value for the team than $\alpha_1$. 

237
of probabilities $\omega_{12}$, $\omega_{13}$, $\omega_{23}$, and $\omega_{123}$. In other words, one can states that the protocol $\alpha_1$ is optimal for the team \{2, 3\}.

Finally, we consider the largest team \{1, 2, 3\} made of all three Musketeers. As in the previous case, one can note that the only relevant alternative protocol to $\alpha_1$ is $\alpha_9$ from Table C.4\footnote{Indeed, it is straightforward to show from Table 6.6 that any other protocol $\gamma$ such that $\gamma^{(1,2,3)} \neq (C, A, B)$ and $\gamma^{(1,2,3)} \neq (C, C, C)$ would always yield a lower expected value for the team than $\alpha_1$.}. Here again, it is straightforward to show that protocol $\alpha_9$ can never strictly outperform $\alpha_1$ for the team \{1, 2, 3\}, that is, $EV_{\{1,2,3\}}(\alpha_1) \geq EV_{\{1,2,3\}}(\alpha_9)$ for any values of probabilities $\omega_{12}$, $\omega_{13}$, $\omega_{23}$, and $\omega_{123}$. In other words, one can states that the protocol $\alpha_1$ is optimal for the team \{1, 2, 3\}.

The first observation one can make from this analysis is that no team reasoner has any interest to deviate from protocol $\alpha_1$. In fact, only the conditions (4) and (5) allow to render $\alpha_1$ non-optimal respectively for Porthos and Aramis when they reason in I-mode (i.e., when they are strictly self-interested).

Furthermore, let us consider the following protocol $\beta_1$:

\[
\begin{align*}
\beta_1^{(2)} &= A \\
\beta_1^{(1)} &= B \\
\beta_1^{(1,2)} &= (A, A) \\
\beta_1^{(1,3)} &= (B, B) \\
\beta_1^{(2,3)} &= (A, B) \\
\beta_1^{(1,2,3)} &= (C, A, B)
\end{align*}
\]

One should note that the symmetrical property in the payoff matrix of the game (i.e., exchanging Athos’s payoffs with Aramis’s would not alter the game) allows us to state that the protocol $\beta_1$ shares the same properties as $\alpha_1$\footnote{The only difference between protocols $\beta_1$ and $\alpha_1$ is that player 1 (i.e., Athos) selects $B$ in $\beta_1$ instead of $A$ in $\alpha_1$ while being self-regarding (i.e., in the I-mode).}, that is, $\beta_1$ is an UTI equilibrium if and only if the two following conditions hold:

\[
\begin{align*}
\omega_{123} &\leq \frac{5}{7} \cdot (1 - \omega_{12}) \\
\omega_{123} &\leq \frac{5}{2} \cdot \omega_{12}
\end{align*}
\]

Therefore, following Conditions (1), (2), (3), (4), (5), (8), and (9), it is straightforward to show that both protocols $\alpha_1$ and $\beta_1$ are always UTI equilibria whenever $\omega_{12} > \omega_{123}$ and $\omega_{13} > \omega_{123}$. In other words, whenever the probability of identifying with the largest group is negligible compared to the probabilities of identifying with some sub-coalition, the corresponding UTI structure cannot yield a unique
solution where all individuals select $C$. In this case, one should note that the study of additional solutions simply becomes irrelevant as the resulting multiple equilibria can only render the model indecisive. As a result of this analysis, it appears that this situation illustrates the disagreement that can exist between Bacharach’s UTI structure and our ST game in terms of the predicted outcome when sub-coalitions are formed.
References


J. Brandts, A. Cabrales, and G. Charness. Forward induction and the excess capacity puzzle: An experimental investigation. 2003. 84


C. Camerer. The promise and success of lab-field generalizability in experimental economics: a critical reply to levitt and list. *Available at SSRN 1977749*, 2011. 42

242


N. Gold and R. Sugden. Theories of team agency. 2007. 94


J. Hintikka. *Knowledge and belief: an introduction to the logic of the two notions*. Cornell University Press Ithaca, 1962. 4, 26, 185


M. Jackson. The economics of social networks. 2005. 180, 196


246
2, 14, 18, 183

2, 14, 18, 183

New York, NY, 2004. 2, 14, 18, 183

24:111–140, 1985. 70

168–188, 2010. 177

M. Rabin. Incorporating fairness into game theory and economics. *American
Economic Review*, 83(5):1281–1302, 1993a. 72, 85, 188

M. Rabin. Incorporating fairness into game theory and economics. *The American
Economic Review*, pages 1281–1302, 1993b. 150

R. Ramanujam and S. Simon. A logical structure for strategies. In *Proc. of LOFT*
7, pages 151–176. Amsterdam University Press, 2008a. 70

R. Ramanujam and S. E. Simon. Dynamic logic on games with structured strate-

139, 140, 170, 172, 189


R.W. Rosenthal. Games of perfect information, predatory pricing and the chain-

243–348, 1995. 37


J.R. Searle. Collective intentions and actions john r. searle. *Intentions in commu-
nication*, page 401, 1990. 146, 194

247

Q. Shahriar. Forward induction works! an experimental study to test the robustness and the power. *Working Papers*, 2009. 84

A. Smerilli. We-thinking and vacillation between frames: filling a gap in bacharach’s theory. *Theory and Decision*, pages 1–22, 2010. 97, 98


248


249
